UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

December 31, 2002

DSSD A.C.E. REVISION II MEMORANDUM SERIES PP#43

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

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

Prepared by: Rosemary Byrne, Michael Beaghen
Mathematical Statisticians,
Decennial Statistical Studies Division

Mary H. Mulry
Principal Researcher
Statistical Research Division

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

Attached is the A.C.E. Revision II Report: Clerical Review of Census Duplicates. Please direct any comments or questions to Rosemary Byrne at 301-763-4251or Michael Beaghen at 301-763-9258.

cc: DSSD A.C.E. REVISION II MEMORANDUM SERIES Distribution List

A.C.E. Revision II Report
#PP-43
December 31, 2002

Clerical Review of Census Duplicates

Rosemary Byrne,
Michael Beaghen, and
Mary Mulry

Decennial Statistical
Studies Division

U S C E N S U S B U R E A U

Helping You Make Informed Decisions

CONTENTS

LIST OF TABLES

EXECUTIVE SUMMARY

What were the goals of the Clerical Review of Census Duplicates (CRCD)?

The main goal of the CRCD was to examine the quality of duplicates identified by the statistical matching component of the Further Study of Person Duplication (FSPD) in Census 2000. The duplicates identified in the FSPD are a key input in the Accuracy and Coverage Evaluation (A.C.E.) Revision II estimation. The CRCD also examined the quality of the Census Administrative Records Duplication Study (CARDS), which was itself an evaluation of the FSPD. Another goal of the CRCD was to determine how many duplicates not detected by the FSPD, were then found either by CARDS or by the CRCD review of other household members within the households with duplicate links.

Were the duplicates identified by the FSPD confirmed by CRCD?

The CRCD confirmed that, generally, the duplicates identified by the FSPD represented duplicates. 94.9% of the E-sample people and 96.3% of P-sample people identified as having duplicate links in the statistical matching part of the FSPD were confirmed by the CRCD.

Did the duplicates falsely identified by the FSPD have a large effect on the A.C.E. Revision II dual system estimates?

The effect of falsely identified duplicates on the dual system estimates was small. There were 47,311 weighted E-sample people and 83,789 weighted P-sample people that were classified as nonduplicates by the CRCD. Additionally, there were 16,336 weighted E-sample and 33,732 weighted P-sample identified as duplicates but classified as unresolved by the CRCD. The overall effect requires consideration of such factors as whether the P-sample nonresidents were matches or nonmatches, and whether the E-sample people were correct or erroneous enumerations, which can be more closely examined in future research. However, the size of the likely effect on the dual systems estimates is just a fraction of the number of false duplicates.

Were the duplicates found by the CARDS but not by the FSPD confirmed by CRCD?

The CRCD determined that about half of the duplicates identified by the CARDS but not by the FSPD did not represent duplication. However, those CARDS duplicates found by the exact matching tended to be confirmed as duplicates. In the E-sample analysis, the CRCD found that of the 1,194,656 weighted people found by the CARDS and not by either FSPD statistical or exact matching, 47.3% were denied and 15.4% were undetermined. In the P-sample analysis the additional CARDS duplicates were even less likely to be confirmed, with 56.4% denied by the CRCD to be duplicates and 15.1% undetermined. CARDS had two phases of matching, one included address information and the other was a name search. The agreement rate between CRCD and CARDS was not assessed for the two phases separately, but is likely to be different.

Were many of the potential duplicates linked but rejected by the FSPD confirmed by CRCD?

The CRCD confirmed that most of the cases linked as potential duplicates but rejected as duplicates by the statistical matching were not duplicates. Only 4.6% of the 3,977,543 weighted E-sample cases and 26.3% of 1,178,059 P-sample cases were confirmed to be duplicates by the CRCD.

Were there many additional duplicates found by CRCD in the households where the statistical matching component of the FSPD already found duplicates?

The CRCD found only a very small number of additional duplicates in households that were already found to have duplicates in the statistical matching component of the FSPD. There were only 46 unweighted E-sample and 73 unweighted P-sample additional duplicates found in the households where FSPD had already found duplicates.

1. BACKGROUND

The primary goal of the Clerical Review of Census Duplicates (CRCD) was for analysts at the National Processing Center (NPC) to examine the quality of duplicates identified by the statistical matching component of the Further Study of Person Duplication (FSPD) in Census 2000. The Accuracy and Coverage Evaluation (A.C.E.) Revision II estimates used estimates of duplicate census enumerations produced by the Further Study of Person Duplication. FSPD used a computer algorithm to identify the duplicates. The analysts reviewed 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). 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.

CRCD reviewed 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.

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 2001) 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 found that the estimate of duplicate census enumerations measured by A.C.E. was less than the estimate from the 1990 Post Enumeration Survey (Jones and Feldpausch 2001).

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. 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 duplication of females 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 methodology similar to that of the ESCAPII Report 20 was 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 performed a national match of E-sample and P-sample records to census enumerations on the Hundred Percent Census Unedited File (HCUF). The algorithm used a statistical matching methodology that assigned a probability of linked records being a match. Links with probabilities above specified thresholds were considered duplicates. The thresholds vaned by the geographical distance between the pair and were set for five 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 differed from the matching for duplication discussed in ESCAPII Report 20, which was exact matching on name and birth date. The statistical matching was augmented by exact matching for the A.C.E. Revision II estimation, but those links were not 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 and Bauder 2002) examined the effectiveness of the FSPD methodology with administrative records through the Census Numident File and the Statistical Administrative Records System 2000 (StARS 2000). CARDS confirmed or denied duplicate links identified by the FSPD. In addition, CARDS identified duplicates missed by FSPD.

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.

2. METHODOLOGY

2.1 Design of matching operation

The analysts reviewed the whole households of duplicates. The analysts had the information for all household members, not just those designated as duplicates. For census enumerations, the information was found on the HCUF. For P-sample cases, the information was collected in the A.C.E. computer-assisted person interview (CAPI) interview.

The analysts entered a code indicating whether a pair is a duplicate along with "why" codes that indicate the reason for declaring the pair a duplicate or denying the duplication and also included notes, if applicable. There were seven categories for the "why" codes: Insufficient Information, Characteristics, Household Composition, Other Residence, Nickname, Duplicate Housing Unit, Other Reason. For the household members that were not designated as having a duplicate by FSPD or CARDS, the analysts entered a code indicating whether a duplicate was found. If there was a 'better' duplicate in the census household other than the one designated by FSPD or CARDS, the analysts recorded a code showing the duplicate was rearranged.

2.2 Sample selection

CRCD reviewed households with duplicates in the Evaluation Sample clusters (Davis & Raglin 2001). The review included duplicates from both the E sample and P sample. Also included were pairs that FSPD linked but did not declare to be duplicates because the probability of being a duplicate was too low. For the E sample, the review was restricted to duplicates between enumerations in the E sample and census enumerations outside the A.C.E. search area (Childers 2001). For the P sample, the review was restricted to households with duplicates between nonmovers and census enumerations outside the search area and does not include links to deleted census enumerations[1].

The review was restricted to households where FSPD finds more than one member was duplicated although households with only CARDS duplicates may have only one. Additional cases from CARDS did not include links to group quarters. The reason for restricting the additional cases to links between housing units was that we believed that few additional duplicates would be found between a household and a group quarters residence. As a result, we had only an estimate of additional duplicates between housing units with the type of duplication included in the study and other housing units. We do not have an estimate of additional duplicates between housing units and group quarters.

The clerical workload included a total of 18,713 links in 11,935 housing units (work units). From the E sample there were 10,248 links in 6,412 housing units while 8,465 links in 5,523 housing units were from the P sample.

2.3 Review of duplicates

[1]P-sample removed persons who linked to census enumerations were inadvertently included in the review sample. Additionally, reinstated and deleted census units were also included among the links. The results for all such cases are presented in separate tables.

The NPC analysts determined whether the sets of two enumerations refer to the same person. The analysts assigned a "why" code that indicates the reason for declaring the pair a duplicate, denying the duplication, or not being able to decide. When the analyst could not decide, the case is considered unresolved. This occurs most often when one or both of the linked records contains an insufficient amount of information to consider the pair a duplicate. In addition to reviewing the linked pairs, the analysts also reviewed household members not linked by FSPD to determine if they too had duplicates.

3. LIMITATIONS

1. The study was restricted to households with two or more duplicates in another housing unit. The study did 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 statistical matching found none.

2. The study can only find missed duplicates within households where duplicate links were identified by the statistical matching component of FSPD and/or CARDS

4. RESULTS

4.1. E-sample Results

The original CRCD plan included only duplicates to census records eligible for the E sample. Duplicates to reinstates and deletes were not included in the original CRCD plan because we believed they were likely to be sound since they were already identified as duplicates or possible duplicates by the census. However, some E-sample records in the CRCD study included census units which were deleted or reinstated, selected through CARDS. For this analysis, we will consider the duplicate links to deleted and reinstated people separately from the other records. Since the duplicates the CRCD reviewed are only those that were identified by CARDS, they are not representative of reinstates and deletes.

Tables 4.1.1 and 4.1.2 show the CRCD coding for each E-sample case (except those linking to deleted and reinstated units) sent for CRCD review. The columns in these tables represent which study, FSPD or CARDS, identified the case as a duplicate and how the other study identified the case. CARDS duplicates are those with a CARDS status of "confirmed", or those listed in the "CARDS only" columns. Cases that FSPD linked but did not call a duplicate are those links determined by FSPD to have a probability of duplication below the threshold to be considered a duplicate. A CARDS status of "denied" means CARDS concluded that the two enumerations were different. A CARDS status of "undetermined" means that CARDS did not have identifying information available for one or both of the enumerations in an FSPD link and therefore could not assess the duplicate status. The "CARDS duplicate" columns show cases identified by CARDS but not identified by the statistical matching component of the FSPD. Some of these cases were also identified during the exact matching part of FSPD. However, these cases are not probability samples of the exact matching, only of cases found by both exact matching and CARDS. Table 4.1.1 shows the unweighted results and 4.1.2 shows the weighted results.

Table 4.1.1 shows that for the unweighted counts, 36.8% of the duplicate links reviewed by analysts were determined to be true duplicates, whereas 57.4% were not. For the weighted results, Table 4.1.2 shows 2.5 million (34.5%) of the records were considered to be duplicates by the analysts while 4.4 million (60.5%) were not.

In Tables 4.1.4 - 4.1.6, we analyze these results further by considering these individually by the three possible FSPD statuses;

- those identified as duplicates by the statistical matching component of FSPD
- those linked by FSPD's statistical matching, but not declared duplicates
- those not identified by FSPD's statistical matching

Table 4.1.1 E-sample Duplication by Study - Unweighted, Standard Errors in Parentheses

Clerical Review status	**Identified in FSPD's Statistical Matching**						**CARDS Duplicate (Not Identified in FSPD's Statistical Matching)**		**Total**
	FSPD duplicate			FSPD linked but not a duplicate			CARDS status		
	CARDS status			CARDS status					
	confirmed	denied	undetermined	confirmed	denied	undetermined	also identified by exact matching	CARDS only	
duplicate	1261 (73)	24 (5)	456 (33)	139 (16)	31 (8)	164 (19)	935 (44)	608 (31)	3,618 (130) 36.8% (0.9)
not a duplicate	6 (2)	48 (9)	26 (9)	40 (7)	3662 (128)	757 (44)	126 (12)	969 (46)	5,634 (172) 57.4% (0.9)
unresolved	17 (6)	0 (0)	10 (4)	10 (4)	42 (8)	49 (9)	162 (16)	278 (19)	568 (33) 5.8% (0.3)
Total	1,284 (74) 13.1% (0.6)	72 (11) 0.7% (0.1)	492 (35) 5.0% (0.3)	189 (18) 1.9% (0.2)	3,735 (129) 38.0% (0.9)	970 (52) 9.9% (0.4)	1,223 (50) 12.4% (0.5)	1,855 (68) 18.9% (0.6)	9,820 (255) 100.00%

Table 4.1.2 E-sample Duplication by Study - Weighted, Standard Errors in Parentheses

Clerical Review status	**Identified in FSPD's Statistical Matching**						**CARDS Duplicate (Not Identified in FSPD's Statistical Matching)**		**Total**
	FSPD duplicate			FSPD linked but not a duplicate			CARDS status		
	CARDS status			CARDS status					
	confirmed	denied	undetermined	confirmed	denied	undetermined	also identified by exact matching	CARDS only	
duplicate	922,325 (59,472)	7,737 (2,101)	262,702 (23,708)	90,092 (14,930)	18,239 (5,895)	76,603 (11,086)	695,968 (36,984)	445,703 (30,309)	2,519,371 (97,056) 34.5% (1.1)
not a duplicate	3,536 (2,378)	35,654 (9,456)	8,121 (3,278)	22,145 (4,911)	3,248,663 (143,023)	459,892 (30,880)	72,647 (9,549)	564,881 (32,568)	4,415,540 (163,478) 60.5% (1.1)
unresolved	10,841 (4,311)	0 (0)	5,496 (2,853)	5,514 (2,214)	30,504 (9,226)	25,890 (7,340)	102,751 (12,495)	184,071 (16,765)	365,067 (25,572) 5.0% (0.3)
Total	936,702 (59,639) 12.8% (0.7)	43,391 (9,679) 0.6% (0.1)	276,320 (24,290) 3.8% (0.3)	117,752 (15,824) 1.6% (0.2)	3,297,406 (143,797) 45.2% (1.1)	562,385 (34,514) 7.7% (0.4)	871,366 (40,427) 11.9% (0.5)	1,194,656 (52,033) 16.4% (0.7)	7,299,977 (209,606) 100%

Because the analysts reviewed all members of the household containing the duplicate links, additional duplicates may have been identified. Table 4.1.3 shows the status for all members of the duplicate households. This shows that of the 23,100 persons in the linked households, 46 household members (0.2%) who were not previously identified as duplicates by FSPD or CARDS were determined to be duplicates by the analysts.

Table 4.1.3 E-sample, Final Disposition Based on Clerical Coding - Weighted and Unweighted, Standard Errors in Parentheses

CRCD classification	**unweighted**	**weighted**
duplicates (CRCD confirmed the links)	3,618 (130)	2,519,371 (97,056)
new clerical duplicates (not previously linked)	46 (8)	n/a
not confirmed asduplicates (CRCD denied or could not confirm the links)	6,202 (183)	4,780,607 (168,855)
non-duplicates (other unlinked household members)	13,234 (451)	n/a
total	23,100 (666)	7,299,977 (209,606)

The following tables (4.1.4 - 4.1.6) show the distribution (weighted) of the clerical coding separately for each possible FSPD outcome.

Table 4.1.4 shows the cases considered to be duplicates by FSPD. Ignoring the CARDS status, the clerical coding confirmed 94.9% of the FSPD duplicates. Of the 3.8% of the FSPD duplicates considered not to be a duplicate by the clerical coding, 75% were also denied in the CARDS study.

Overall, both CARDS and CRCD agreed that 73.4% (922,325 out of 1.25 million) of the duplicate links found by FSPD's statistical matching were duplicates.

Note that Table 4.1.4 shows an evaluation of 1.25 million of the 3.4 million duplicates found outside the A.C.E. surrounding blocks in Table 2 in the "Further Study of Person Duplicates" (Mule, 2002).

Table 4.1.4 E-sample Duplication by Study, FSPD Statistical Matching Duplicate - Weighted, Standard Errors in Parentheses

Clerical Review status	**Identified in FSPD's Statistical Matching**			**Total**
	FSPD duplicate			
	CARDS status			
	confirmed	denied	undetermined	
duplicate	922,325 (59,472)	7,737 (2,101)	262,702 (23,708)	1,192,765 (71,834) 94.9% (1.0)
not a duplicate	3,536 (2,378)	35,654 (9,456)	8,121 (3,278)	47,311 (11,223) 3.8% (0.9)
unresolved	10,841 (4,311)	0.0 (0.0)	5,496 (2,853)	16,336 (6,602) 1.3% (0.5)
Total	936,702 (59,639) 74.5% (1.6)	43,391 (9,679) 3.5% (0.8)	276,320 (24,290) 22.0% (1.5)	1,256,413 (73,671) 100%

Table 4.1.5 shows the clerical coding for the cases linked by FSPD but considered to be below the threshold to be considered a duplicate. Here we see that disregarding the CARDS status, 93.8% of these links were also not considered to be a duplicate by the analysts. The CRCD did determine that 4.6% of the links not considered to be duplicates by FSPD were indeed duplicates. About half of these (90,092) were also identified as duplicates by CARDS.

Overall, both CARDS and CRCD determined that 81.7% (3.2 million out of 3.9 million) of the links FSPD found but did not declare duplicates were not duplicates.

Table 4.1.5 E-sample Duplication by Study, FSPD linked but not a duplicate - Weighted, Standard Errors in Parentheses

Clerical Review status	**Identified in FSPD's Statistical Matching**			**Total**
	FSPD linked but not a duplicate			
	CARDS status			
	confirmed	denied	undetermined	
duplicate	90,092 (14,930)	18,239 (5,895)	76,603 (11,086)	184,934 (21,891) 4.6% (0.5)
not a duplicate	22,145 (4,911)	3,248,663 (143,023)	459,892 (30,880)	3,730,701 (153,928) 93.8% (0.6)
unresolved	5,514 (2,214)	30,504 (9,226)	25,890 (7,341)	61,908 (12,331) 1.6% (0.3)
Total	117,752 (15,824) 3.0% (0.4)	3,297,406 (143,797) 82.9% (0.9)	562,385 (34,514) 14.1% (0.8)	3,977,543 (157,888) 100%

Table 4.1.6 shows the clerical coding results for cases identified by CARDS, but not by the statistical matching part of FSPD. This table includes cases identified by the exact matching part of FSPD but not identified in the statistical matching part of FSPD. Rather, they were among the CARDS cases yet happened to be linked in the exact matching.

For cases identified by CARDS but not by FSPD's statistical matching component, the confirmed duplication rate from the clerical matching is much lower, 55.3%. Of these, 61% (695,968 out of 1.14 million) were also identified by the exact matching.

Note that about 175,398 (20%) of the 871,366.3 links also identified by exact matching were not considered to be duplicates by CRCD. However, no conclusions can be drawn about the overall quality of exact matching based on these results since these were not sampled for the CRCD review.

Table 4.1.6 E-sample Duplication by Study, CARDs Only - Weighted, Standard Errors in Parentheses

Clerical Review status	**CARDS status**		**Total**
	also identified by exact matching	CARDS only	
duplicate	695,968 (36,984)	445,703 (30,309)	1,141,672 (51,642) 55.3% (1.6)
not a duplicate	72,647 (9,549)	564,881 (32,568)	637,528 (35,301) 30.9% (1.4)
unresolved	102,751 (12,495)	184,071 (16,765)	286,822 (21,165) 13.9% (0.9)
Total	871,366 (40,427) 42.2% (1.4)	1,194,656 (52,033) 57.8% (1.4)	2,066,022 (71,515) 100%

Next we considered the results of the clerical coding by "why" code for the CARDS only duplicate links. The analysts assigned a "why" code that indicates the reason for declaring the pair a duplicate, denying the duplication, or not being able to decide. There were seven categories for the "why" codes:

- Insufficient Information - the data available for the pair was insufficient or incomplete and no determination could be made
- Characteristics - the decision was based on demographic information for the linked pair
- Household Composition - the decision was based on the composition of the household
- Other Residence - the decision was based on knowledge of another residence for the person
- Nickname - the name is a common nickname and is likely to be the same
- Duplicate Housing Unit - the housing unit is a duplicate
- Other Reason

Table 4.1.7 shows the "why" codes for the CARDS only cases. While "household composition" was not used as frequently to confirm the duplicates, it was used often to deny the links.

For links considered to be duplicates by CRCD, the proportion of cases with a why code of "characteristics" is about four times the number coded "household composition" for both the CARDS only cases and those also identified by exact matching. However, for the links CRCD classified as non-duplicates, the exact matching "why" code of "household composition" is more than ten times more prevalent than "characteristics". For the CARDS only cases, "household composition" is about twice as prevalent. This difference is likely due to the fact that the exact matching does not incorporate household composition into its process.

Table 4.1.7 E-sample CARDS only, Why Codes - Weighted, Standard Errors in Parentheses

Clerical Review "Why" Codes		CARDS status		Total
		also identified by exact matching	CARDS only	
duplicate	household composition	59,687 (10,446)	47,899 (8,676)	107,585 (14,421) 5.2% (0.7)
	characteristics	366,604 (25,257)	205,507 (18,657)	572,111 (32,313) 27.7% (1.3)
	other residence	269,678 (24,600)	192,275 (21,671)	461,952 (36,677) 22.4% (1.5)
	rearranged duplicate	0.0 (0.0)	23 (23)	23 (23) 0.0% (0.0)
not a duplicate	household composition	67,588 (9,295)	367,704 (26,758)	435,292 (29,703) 21.1% (1.2)
	characteristics	5,059 (2,284)	197,178 (16,939)	202,236 (17,114) 9.8% (0.8)
unresolved	insufficient information	102,751 (12,495)	184,071 (16,765)	286,822 (21,165) 13.9% (0.9)
Total		871,366 (40,427) 42.2% (1.4)	1,194,656 (52,033) 57.8% (1.4)	2,066,022 (71,515) 100%

4.2. E-sample Results for Deleted and Reinstated Units

The CARDS only links also included links to deleted or reinstated units. These were not included in the tables in section 4.1, but are tabulated here. We expect the duplication rate to be higher for these cases since the housing units have previously been identified as duplicates or possibly duplicates. Overall, 91% of these cases were confirmed duplicates. For the cases also identified by exact matching, 98.5% were considered duplicates by the analysts.

Table 4.2.1 E-sample Cases Linked to Reinstates and Deletes, Duplication by Study - Weighted, Standard Errors in Parentheses

Clerical Review status	**CARDS Duplicate (Not Identified in FSPD's Statistical Matching)**		**Total**
	CARDS status		
	also identified by exact matching	CARDS only	
duplicate	556,825 (62,457)	24,956 (9,266)	581,781 (63,175) 91.0% (2.0)
not a duplicate	5,982 (4,172)	36,833 (9,764)	42,814 (10,605) 6.7% (1.7)
unresolved	2,720 (1,506)	11,907 (6,222)	14,627 (6,399) 2.3% (1.0)
Total	565,527 (62,566.) 88.5% (2.3)	73,696 (14,765) 11.5% (2.3)	639,223 (64,807) 100%

Table 4.2.2 shows the disposition of all persons in the E-sample households which had duplicate links to reinstated and deleted units. The CRCD identified 64 new duplicates, approximately 9% of the household members. Recall that for the rest of the E-sample cases that linked to E-sample eligible cases, Table 4.1.3 showed 0.2% new duplicates.

Table 4.2.2 E-sample Cases Linked to Reinstates and Deletes, Final Disposition Based on Clerical Coding, Weighted and Unweighted, Standard Errors in Parentheses

CRCD classification	**unweighted**	**weighted**
duplicates (CRCD confirmed the links)	388 (34)	581,781 (63,175)
new clerical duplicates (not previously linked)	64 (10)	n/a
not confirmed as duplicates (CRCD denied or could not confirm the links)	40 (6)	57,441 (12,354)
non-duplicates (other unlinked household members)	224 (29)	n/a
Total	716 (56)	639,223 (64,807)

4.3 P-sample Results for Nonmovers

The P-sample results for P-sample nonmovers are presented in Tables 4.3.1 - 4.3.4. We present the overall status for the unweighted totals in Table 4.3.1, the totals weighted by the residence probability in Table 4.3.2, and the totals weighted by the probability of nonresidence in Table 4.3.3. Lastly, Table 4.3.4, shows the final disposition for all P-sample nonmovers in the households of the linked nonmovers.

Overall, the CRCD classified 66.2% of the unweighted P-sample nonmover duplicate links as confirmed duplicates. This is almost twice the percent confirmed in the E-sample (see Table 4.1.1).

This lower overall confirmed rate of the E-sample links is due to the large number, about four million, of E-sample links identified as potential duplicates in the statistical matching but rejected. Since CRCD denied that most of these links were duplicates the difference is not a cause of concern.

We break the weighted P-sample nonmovers into two tables, 4.3.2 for residents and 4.3.3 for nonresidents. Table 4.3.2 is more relevant to the A.C.E. Revision II because it shows links to people who figured in the A.C.E. Revision II estimation. In Table 4.3.3 the CRCD classified 88.8% of the links to non-residents as confirmed, higher than the 66.2% for residents. This difference is likewise attributable to differences in the population of duplicates compared rather than to the duplicate search itself. P-sample people who are truly duplicated are often non-residents.

In Table 4.3.2 note that many of the patterns we saw in the E-sample tables are present for the P-sample nonmovers.

- The number of duplicates identified in the statistical matching but denied by the CRCD is modest, 83,781.8, with another 33,732.3 undetermined.
- The proportion of statistical matches that were confirmed, 96.3%, is comparable to that of the E-sample.
- The number of potential links rejected by the statistical matching that were determined to be duplicates by CRCD is 309,622.5 (26.3%) out of 1,178059, which is proportionally larger than the 4.6% we saw in the E-sample.
- In the P-sample analysis the additional CARDS duplicates were even less likely to be confirmed than in the E-sample, with 56.4% denied by the CRCD to be duplicates and 15.1% undetermined.

In Table 4.3.3 the proportions of confirmed among the nonresidents are across the board higher than what we saw among the residents in 4.3.3. In Table 4.3.4, as in the E-sample, we see only a modest number, 73 (0.4%), of additional duplicates found by the analysts. This is twice as many as were found in the E-sample (see Table 4.1.3).

Table 4.3.1 P-sample Nonmovers, Duplication by Study - Unweighted, Standard Errors in Parentheses

Clerical Review status	**Identified in FSPD's Statistical Matching**						**CARDS Duplicate (Not Identified in FSPD's Statistical Matching)**		**Total**
	FSPD duplicate			FSPD linked but not a duplicate			CARDS status		
	CARDS status			CARDS status					
	confirmed	denied	undetermined	confirmed	denied	undetermined	also identified by exact matching	CARDS only	
duplicate	2,155 (163)	40 (7)	607 (54)	171 (18)	16 (6)	97 (13)	960 (43)	528 (27)	4,574 (235) 66.2% (1.3)
not a duplicate	2 (1)	44 (9)	22 (8)	59 (9)	274 (23)	179 (19)	233 (17)	973 (44)	1,786 (72) 25.9% (1.1)
unresolved	10 (3)	1 (1)	15 (5)	19 (5)	14 (4)	41 (9)	187 (16)	257 (18)	544 (33) 7.9% (0.5)
Total	2,167 (163) 31.4% (1.5)	85 (11) 1.2% (0.2)	644 (55) 9.3% (0.6)	249 (22) 3.6% (0.3)	304 (25) 4.4% (0.4)	317 (25) 4.6% (0.4)	1,380 (51) 20.0% (0.7)	1,758 (61) 25.5% (0.9)	6,904 (264) 100%

Table 4.3.2 P-sample Nonmovers, Duplication by Study- Weighted Residents, Standard Errors in Parentheses

Clerical Review status	**Identified in FSPD's Statistical Matching**						**CARDS Duplicate (Not Identified in FSPD's Statistical Matching)**		**Total**
	FSPD duplicate			FSPD linked but not a duplicate			CARDS status		
	CARDS status			CARDS status					
	confirmed	denied	undetermined	confirmed	denied	undetermined	also identified by exact matching	CARDS only	
duplicate	2,501,879 (213,787)	39,659 (10,385)	544,882 (56,451)	195,438 (30,455)	23,608 (11,446)	90,577 (17,143)	1,267,247 (74,660)	648,382 (46,360)	5,311,671 (273,946) 62.2% (1.5)
not a duplicate	2,762 (1,985)	59,786 (19,057)	21,234 (10,439)	73,670 (13,745)	507,531 (59,091)	200,025 (29,167)	332,983 (31,995)	1,285,066 (71,626)	2,483,056 (122,158) 29.1% (1.3)
unresolved	19,780 (9,620)	0.0 (0.0)	13,952 (7,722)	27,522 (11,781)	30,543 (13,418)	29,147 (8,657)	284,712 (36,440)	344,021 (33,337)	749,678 (58,310) 8.8% (0.7)
Total	2,524,421 (213,891) 29.5% (1.8)	99,445 (21,631) 1.2% (0.2)	580,067 (57,884) 6.8% (0.6)	296,630 (35,826) 3.5% (0.4)	561,682 (61,449) 6.6% (0.7)	319,748 (34,611) 3.7% (0.4)	1,884,942 (91,235) 22.1% (1.0)	2,277,469 (97,850) 26.5% (1.1)	8,544,404 (323,510) 100%

Table 4.3.3 P-sample Nonmovers, Duplication by Study - Weighted Non-residents, Standard Errors in Parentheses

Clerical Review status	**Identified in FSPD's Statistical Matching**						**CARDS Duplicate (Not Identified in FSPD's Statistical Matching)**		**Total**
	FSPD duplicate			FSPD linked but not a duplicate			CARDS status		
	CARDS status			CARDS status					
	confirmed	denied	undetermined	confirmed	denied	undetermined	also identified by exact matching	CARDS only	
duplicate	232,099 (71,794)	851 (592)	69,810 (18,924)	8,782 (3,844)	273 (273)	15,663 (8,654)	336,285 (99,772)	168,099 (52,464)	831,864 (213,941) 88.8% (3.6)
not a duplicate	0.0 (0.0)	165 (146)	2,341 (1,904)	0.0 (0.0)	23 (18)	8,162 (2,562)	3,367 (1,581)	40,449 (10,730)	54,507 (11,294) 5.8% (2.0)
unresolved	335 (255)	0.0 (0.0)	8,464 (7,055)	0.0 (0.0)	1,126 (1,126)	2,792 (1,948)	7,470 (4,784)	29,733 (11,167)	49,921 (14,201) 5.3% (2.1)
Total	232,435 (71,793) 24.8% (3.4)	1,016 (610) 0.1% (0.1)	80,615 (20,261) 8.6% (3.1)	8,782 (3,844) 1.0% (0.5)	1,422 (1,159) 0.2% (0.1)	26,618 (9,220) 2.8% (1.3)	347,123 (99,878) 37.1% (4.0)	238,282 (54,644) 25.5% (2.7)	936,291 (215,072) 100%

Table 4.3.4 P-sample Nonmovers, Final Disposition Based on Clerical Coding - Weighted and Unweighted, Standard Errors in Parentheses

CRCD classification	**unweighted**	**weighted**
duplicates (CRCD confirmed the links)	4,574 (235)	6,143,535 (352,162)
new clerical duplicates (not previously linked)	73 (12)	n/a
not confirmed asduplicates (CRCD denied or could not confirm the links)	10,701 (376.3)	3,337,161 (148,383)
non-duplicates (other unlinked household members)	2,330 (89)	n/a
Total	17,678 (553) 100%	9,480,695 (393,491) 100%

4.4 P-sample Results for Movers and Removed, and Nonmovers Linked to Reinstated or Deleted Units

In this section, we present results for the P-sample duplicates where the P-sample person was a mover or was removed based on whether the person stayed in group quarters or another residence. Additionally, this section includes the results for the P-sample duplicate links to the reinstated or deleted units.

Table 4.4.1 shows the weighted totals of the CRCD coding for the P-sample movers and removed persons. We expect a very high duplication rate since these people have recently moved or have indicated that they reside elsewhere. Overall, the rate of confirmed duplication for this group is 93.2%.

Table 4.4.1 P-sample Movers and Removed, Duplication by Study - Weighted, Standard Errors in Parentheses

Clerical Review status	**Identified in FSPD's Statistical Matching**						**CARDS Duplicate (Not Identified in FSPD's Statistical Matching)**		**Total**
	FSPD duplicate			FSPD linked but not a duplicate			CARDS status		
	CARDS status			CARDS status					
	confirmed	denied	undetermined	confirmed	denied	undetermined	also identified by exact matching	CARDS only	
duplicate	846,493 (114,957)	5,091 (3,605)	127,805 (22,419)	122,897 (32,844)	2,441 (2,007)	48,022 (19,484)	614,664 (60,265)	312,770 (35,015)	2,080,181 (188,023) 93.2% (1.2)
not a duplicate	0.0 (0.0)	3,859 (3,859)	0.0 (0.0)	0.0 (0.0)	7,145 (6,955)	1,305 (1,305)	8,113 (3,963)	62,396 (14,921)	82,819 (17,484) 3.7% (0.8)
unresolved	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	546 (546)	0.0 (0.0)	546 (546)	13,243 (6,539)	54,385 (14,001)	68,722 (15,952) 3.1% (0.7)
Total	846,493 (114,957) 37.9% (3.0)	8,950 (5,279) 0.4% (0.2)	127,805 (22,419) 5.7% (0.9)	123,443 (32,848) 5.5% (1.2)	9,587 (7,237) 0.4% (0.3)	49,873 (19,619) 2.2% (0.8)	636,020 (61,382) 28.5% (2.5)	429,551 (42,881) 19.2% (2.1)	2,231,721 (192,644) 100%

Similarly, we expect the rate of confirmed duplication by CRCD coding to be higher for those persons linked to deleted or reinstated units. Table 4.4.2 shows 85.9% overall confirmed duplication by CRCD coding.

Table 4.4.2 P-sample Nonmover Residents Linked to Deleted and Reinstated Units, Duplication by Study - Weighted, Standard Errors in Parentheses

Clerical Review status	**CARDS Duplicate (Not Identified in FSPD's Statistical Matching)**		**Total**
	CARDS status		
	also identified by exact matching	CARDS only	
duplicate	270,712 (42,172)	39,384 (19,426)	310,096 (46,367) 85.9% (3.5)
not a duplicate	2,571 (1,502)	39,336 (11,113)	41,908 (11,208) 11.6% (3.2)
unresolved	783 (644)	8,016 (5,240)	8,799 (5,279) 2.4% (1.5)
Total	274,067 (42,189) 76.0% (5.8)	86,736 (22,923) 24.0% (5.8)	360,803 (47,784) 100%

Table 4.4.3 shows the status for all members of the duplicate households reviewed by the analysts. This shows that of the 2,395 persons in the linked households, 82 household members (3.4%) who were not previously identified as duplicates by FSPD or CARDS were determined to be duplicates by the analysts.

Table 4.4.3 P-sample Movers, Removed, and Nonmovers Linked to Deleted and Reinstated Units, Final Disposition Based on Clerical Coding - Weighted and Unweighted, Standard Errors in Parentheses

CRCD classification	Movers and Removed		Nonmovers Linked to Reinstated and Deleted Units	
	unweighted	weighted	unweighted	weighted
duplicates (CRCD confirmed the links)	1,203 (97)	2,080,181 (188,023)	216 (32)	310,096 (46,367)
new clerical duplicates (not previously linked)	82 (12)	n/a	47 (12)	n/a
not confirmed as duplicates (CRCD denied or could not confirm the links)	1,026 (72)	151,539 (26,507)	252 (32)	50,707 (12,367)
non-duplicates (other unlinked household members)	84 (11)	n/a	34 (6)	n/a
Total	2,395 (151)	2,231,721 (192,644)	549 (57)	360,803 (47,784)

5. FUTURE RESEARCH

This section discusses topics that would be fruitful for further investigation.

Research that can be conducted with the current data includes:

- How do the accuracy of the FSPD and CARDS compare by geographical distance? That is, how does the quality vary when the duplicate pairs are within the county but the surrounding blocks, within the same state, and in different states?
- How does the accuracy of the FSPD and CARDS compare among demographic characteristics such as race/ethnicity, age, and sex?
- What is the overall estimate of duplication in the Census 2000 determined by combining results from the FSPD, CARDS, and CRCD?
- What are the effects on the A.C.E. Revision II dual system estimates of the falsely identified duplicates?
- What was the effect on FSPD error of the cutoff values chosen in the statistical matching component of FSPD? Could better cutoff values have been chosen?
- Did the quality of duplicates identified by CARDS vary by whether they were identified in the phase that included address information or in the name search phase?

Research that would require additional clerical review or automated processing includes:

- What is the quality of the duplicates identified in the exact matching component of the FSPD? (Currently only those exact matches which were identified by CARDS were in the CRCD). This work involves more clerical review.
- When the CRCD and CARDS disagree as to whether two person records refer to the same person, which is correct? Making this determination could require interviewing the disputed linking people.

6. CONCLUSIONS

The CRCD represents the first time that skilled analysts have clerically reviewed in a systematic way a sizeable sample of duplicate links found by the automated duplicate searches, the FSPD and the CARDS. This review yielded several important insights.

- The most general conclusion from the CRCD is that the links identified by the statistical matching component of the FSPD appear to be genuine duplication.
- The level of false duplication in the FSPD is modest and does not threaten the integrity of the A.C.E. Revision II estimates.
- The number of CRCD confirmed CARDS-only duplicates is large enough to indicate that the FSPD does not find all duplicates that can be identified. At the same time, the large proportion of CARDS-only links that were denied by the CRCD indicates that separate evaluations of the effectiveness of the two phases of matching used by CARDS is necessary.

Lastly we point out that there is still much that can be learned about duplication in Census 2000 both from further analysis of CRCD data and from new studies on census duplication.

7. REFERENCES

Bean, Susanne L. and Bauder, D. Mark (2002). "Census and Administrative Records Duplication Study," DSSD Revised A.C.E. Estimates Memorandum Series #PP-44. Census Bureau, Washington, DC.

Childers, Danny (2001). "Accuracy and Coverage Evaluation: The Design Document," DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1, Revised.

Davis, Mary & Raglin, David (2001). "Creation of Master Data Variance Files for Coverage Measurement Evaluations," Planning, Research and Evaluation Division TXE/2010 Memorandum Series: CM-GES-S-01-R.

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.

Mule, Thomas (2002) "Further Study of Person Duplicates". DSSD Revised A.C.E. Estimates Memorandum Series #PP-51. Census Bureau, 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.

Mulry, Mary (2002). "Chapter 7: Assessing the Estimates," in "Revised ACE: Design and Methodology." Revised A.C.E. Estimates Memorandum Series #PP- 30. Census Bureau, Washington, DC.