

# Person Duplication in the Search Area Measured by the 2000 Accuracy and Coverage Evaluation

## **FINAL REPORT**

This evaluation study reports the results of research and analysis undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation (TXE) Program, designed to assess Census 2000 and inform 2010 Census planning. Findings from the Census 2000 TXE Program reports are integrated into topic reports that provide context and background for broader interpretation of results.

John Jones

---

Decennial Statistical  
Studies Division

## CONTENTS

EXECUTIVE SUMMARY.....	iv
1. BACKGROUND.....	1
2. METHODS.....	3
3. LIMITS.....	5
4. RESULTS.....	6
4.1 How did census person duplication in Census 2000 compare with census person duplication in the 1990 Census?.....	6
4.2 What was the frequency of census person duplication? Did this frequency vary within important variables? What attributes did person duplicates have in common with those that they duplicated? On which attributes did person duplicate pairs disagree?.....	7
4.3 What kinds of housing units contain person duplicates? To what extent did person duplication reflect housing unit duplication? Were there other census operations contributing to census person duplication?.....	14
4.4 What was the frequency of P-sample person duplication? Did this frequency vary within important P-sample person variables?.....	20
5. CONCLUSIONS AND RECOMMENDATIONS.....	23
6. REFERENCES.....	25
7. APPENDIX.....	26

## LIST OF TABLES

Table 1: E-sample Overall Percent Person Duplication.....	7
Table 2: Weighted Person Duplication Percentages by Regional Office (E-sample).....	8
Table 3: Weighted Person Duplication Percentages by Size of Metropolitan Area (E-sample).....	9
Table 4: Weighted Person Duplication Percentages by Type of Census Return (E-sample).....	9
Table 5: Cross Classification of Duplicate E-sample Pairs by Return Type.....	10
Table 6: Weighted Person Duplication Percentages by Domain (E-sample).....	11
Table 7: Weighted Person Duplication Percentages by Age/Sex (E-sample).....	12
Table 8: Cross Classification of Duplicate E-sample Pairs by Age category.....	13
Table 9: Cross Classification of Duplicate E-sample Pairs by Gender.....	13
Table 10: Weighted Person Duplication Percentages by Housing Tenure (E-sample).....	14
Table 11: Cross Classification of Duplicate E-sample Pairs by Housing Tenure.....	14
Table 12: Weighted Person Duplication Percentages by Housing Unit Enumeration Status (E-sample).....	15
Table 13: Weighted Person Duplication Percentages by Type of Address (E-sample).....	16
Table 13a: Weighted Person Duplication Percentages by Type of Small Multiunit.....	16
Table 14: Weighted Person Duplication Percentages by Percentage of Mobile Homes in Cluster (E-sample).....	17

Table 15: Weighted Person Duplication Percentages by Source of Census Address (E-sample).....	18
Table 15a: Weighted Person Duplication Percentage of Source of Census Address that were Primaries to E-sample Person Duplicates.....	19
Table 16: Weighted P-sample Person Duplication Percentages by Region.....	20
Table 17: Weighted P-sample Person Duplication Percentages by Type of Enumeration Area...	20
Table 18: Weighted P-sample Person Duplication Percentages by Type of Basic Address.....	21
Table 19: Weighted P-sample Duplication Percentages by Gender.....	21
Table 20: Weighted P-sample Person Duplication Percentages by Age.....	22
Table 21: Weighted P-sample Duplication Percentages by Housing Tenure.....	22
Table 22: Weighted P-sample Duplication Percentages by Proxy Status.....	22

## EXECUTIVE SUMMARY

This paper examines the Census 2000 duplication as measured by the 2000 Accuracy and Coverage Evaluation. It also examines person duplication within the Accuracy and Coverage Evaluation sample. The Accuracy and Coverage Evaluation was an operation undertaken to evaluate the coverage of Census 2000. It was comprised of the matching of an independent enumeration in a stratified sample of census block clusters against the Census 2000 enumerations in those block clusters. The 2000 Accuracy and Coverage Evaluation included an initial housing unit phase, where housing units in the sampled block clusters were matched against units listed in the January 2000 Decennial Master Address File in those same clusters; a person interview phase, where demographic information was collected from Census Day residents of housing units in the sampled block clusters; and a person match phase, where persons listed in the independent enumeration were matched against the census record of persons in those same clusters.

The results of the 2000 Accuracy and Coverage Evaluation were used to calculate dual system population estimates. The dual system estimator applies a factor that consists of the ratio of the correct enumeration rate to the match rate to the number of data defined census enumerations. The match rate was obtained from the matching of the independent enumeration (**P-sample**) to the Census 2000 enumeration (**E-sample**). The Accuracy and Coverage Evaluation classified each census enumeration as either correct or erroneous, and thereby provided the rate of correct enumeration. Person records identified as duplicates were classified as erroneous enumerations and therefore lowered the correct enumeration rate.

The following limitations apply:

- **The duplicate search was not performed outside of the duplicate search area.** For most clusters, the search area was the block cluster. The remaining clusters were subject to extended search so that the search area included one ring of blocks surrounding the block cluster. This search was also not performed within group quarters of the sampled block clusters. There is another Census 2000 evaluation that reports the outcome of a person duplicate search outside of the duplicate search area. See *Mule, Tom (2001). "ESCAP II: Person Duplication in Census 2000", Executive Steering Committee for A.C.E. Policy (ESCAP II), Report 20, October 11, 2001, U.S. Census Bureau.*
- **The universe of census housing units did not include housing units deleted and later reinstated by a census operation designed to remove potential duplicates.** Therefore, persons enumerated in these units are outside of the scope of this analysis.
- **Tables in this study are based upon Accuracy and Coverage Evaluation production data only.** Person duplicates found by other means are outside of the scope of this study.

The key findings of this study are as follows:

- **Racial and ethnic groups that are traditionally undercounted are also more frequently duplicated.** African Americans and Hispanics have the highest duplication percentages. Because person duplicates are erroneous enumerations, they actually reduce the census undercount.
- **Duplication is more frequent in certain areas of the country.** The New York and Boston regional offices have significantly more census duplication than the rest of the nation. Census duplication is more frequent in large metropolitan areas and in rural areas. P-sample person duplication is significantly less frequent in the Midwest.
- **Census duplication is more prevalent in multi-unit structures.** Person duplication is most frequent in smaller multi-unit structures. It is also more frequent among renters than among owners.
- **Census person duplication is more prevalent in duplicate housing units.** However, most persons enumerated in these units were not coded duplicate.
- **Census person duplication appears to be less frequent in housing units that were in the census inventory in 1990.** It appears to be more frequent in housing units that were added to the census more recently.
- **P-sample person duplication occurred less frequently than E-sample person duplication.** However, there were some qualitative similarities between the two.

The following recommendations stem from the conclusions of this study:

- **It is beneficial to conduct an operation to unduplicate persons and a separate though related operation to unduplicate housing units.**
- **Future efforts to unduplicate persons should emphasize those living in small multiunits and mobile homes.**

## 1. BACKGROUND

The person match phase of the Accuracy and Coverage Evaluation (A.C.E.) began after the census and the A.C.E. person interview phase were completed. Information on persons independently enumerated came from the person interview; these persons are also known as P-sample persons. The persons enumerated in the sampled clusters by Census 2000 were divided into three groups based upon the outcome of a subsampling within large clusters and upon the selection of a subset of clusters for targeted extended search. In person matching, match and residence codes were assigned to P-sample persons and match and enumeration codes were assigned to census persons. Also, a duplicate search was performed for census persons. This search occurred within housing units in the search area and not within group quarters. We are concerned with the level of duplication of these census persons.

The E-sample identification placed census persons into one of the following categories:

- **E-sample persons:** These are persons enumerated in small and medium sized block clusters and persons enumerated in large block clusters that are still in sample after within large block subsampling.
- **Non E-sample persons:** These are persons that are out of sample after within large block subsampling.
- **Surrounding block persons:** These are persons enumerated in the surrounding blocks of clusters chosen for targeted extended search.

A census person record is said to be the duplicate of another census person record if the pair of records refer to the same person. The characteristics used to identify duplicates were name, age, gender, race, Hispanic origin, and street address. The duplicate search was restricted to duplicates of E-sample persons. When two or more persons referred to the same person, all but one of them were coded as duplicates. Each E-sample person that was coded duplicate counted as one erroneous enumeration. When one or more non E-sample persons duplicate an E-sample person, the person who was duplicated counted as less than one erroneous enumeration with the exact fraction depending on the number of non E-sample persons that were duplicates.

Duplicates were linked to the people that they duplicate. Some of the tables in Section 4 were created using a database of linked duplicate pairs. It was possible for census persons to have more than one duplicate; when this happened a separate record was created for each duplicate pair. Duplicated census persons are also known as primaries. There were three types of duplicate pair linkages. They were:

- **E-sample duplicates of E-sample persons:** This was the most common type of duplicate pair. The duplicated (primary) person was either matched to a P-sample person; not matched to the P sample, but correctly enumerated in the Census; or not matched to

the P-sample with unresolved enumeration status. Persons who were erroneously enumerated were not included in duplicate searches.

- **Non E-sample duplicates of E-sample persons:** This was the next most common type of duplicate pair. The duplicated (primary) E-sample person counted as a partial erroneous enumeration.
- **E-sample duplicates of surrounding block persons:** This occurred in clusters chosen for targeted extended search because of errors identified in the initial housing unit phase.

Person duplicates within the P-sample were identified by a search within the P-sample listings. This search was entirely separate from the within census duplicate search. The P-sample person duplication was the outcome of Accuracy and Coverage Evaluation operations only. Estimates of P-sample person duplication are based on data used in person matching production. Like the census, when two or more P-sample enumerations referred to the same person, all but one of them were coded as duplicates. The primary person was included in the match to the census. Unlike the census, P-sample duplicates were simply removed from the P-sample and no further accounting for them was done.

## 2. METHODS

**Table 1** gives the overall weighted percentage of E-sample persons that were duplicates while **Tables 2, 3, 4, 6, 7, 10, 12, 13, 13a, 14 and 15** give this percentage and the associated standard error for each level of the following variables:

- Regional office
- Size of Metropolitan Area
- Type of Census Return
- Racial/Ethnic Domain
- Age/Sex Category
- Tenure
- Housing Unit Enumeration Status
- Type of Structure (Number of Units at Basic Street Address)
- Percentage of Mobile Homes in Cluster
- Source of Address

In these tables both E-sample persons coded duplicate, and E-sample primary (duplicated) persons with links to Non E-sample persons were counted as duplicates. Table 13 gives person duplication percentages by Type of Structure where the levels of Type of Structure are single, small multiunit (2-9 units at BSA) and large multiunits. Table 13a gives person duplication percentages by small multiunit status where the levels are 2 units at BSA and 3-9 units at BSA. **Table 15a** gives the percentage of E-sample persons that were E-sample primaries (correctly enumerated persons with duplicate links) by Source of Address. This table excludes those primaries with duplicate links to Non E-sample persons.

**Tables 16 through 22** give the overall weighted percentage of P-sample persons that were coded duplicate and the associated standard error for each level of the following variables:

- Region
- Type of Enumeration Area
- Type of Basic Address

- Gender
- Age
- Housing Tenure
- Proxy Status

The percentage of duplication is the ratio of the weighted number of duplicates to the weighted number of persons multiplied by 100. In the E-sample, both units with final match code as duplicate and units with duplicate links to non E-sample people are counted as duplicates. People with final match code as duplicate are counted as one erroneous enumeration while units with duplicate links to non E-sample people are counted as a partial erroneous enumeration, with the exact fraction depending upon the number of non E-sample duplicate links (see the Appendix). Only P-sample persons coded as duplicate were counted as duplicates. The person weights reflect the probability of selection in all phases of sampling and the probability of erroneous enumeration. Standard errors of these rates were calculated using stratified jackknife methods by the software package VPLX. The VPLX package uses replication methods to calculate variances of estimates derived from complex surveys as described in Fay (1990). Once these rates and their standard errors are determined, within variable comparisons are made to check for significant differences in the frequency of duplication. These comparisons are made using critical values of t-statistics. These critical values are determined using a multiple comparison of means technique with a Bonferroni adjustment, as described in Hocking (1986). The overall significance level is 10 percent.

**Tables 5, 8, 9, and 11** utilize a database of linked duplicate pairs. If an E-sample person had  $n$  duplicates then the database had  $n$  separate records. Each record of the database contains person characteristics of each member of the linked duplicate pair. Each pair consists of a duplicated person called the primary and the duplicate person. Primary persons are usually matched to P-sample persons or are otherwise correctly enumerated. The duplicate person was coded duplicate and is consequently an erroneous enumeration. The database was used to investigate the agreement on these characteristics of the linked pairs. Duplicate pairs are said to agree if they have identical characteristics. If either or both members of the pair have a missing characteristic, then the pair cannot agree on that characteristic.

### 3. LIMITS

One limitation of this study concerns the available universe of eligible housing units. Before the beginning of the Accuracy and Coverage Evaluation, the census flagged housing units it thought to be potential duplicates. Some of these flagged units were later deleted while some were kept in the census. None of these units were in the Accuracy and Coverage Evaluation housing unit universe. Therefore, people enumerated in housing units flagged by the census were not included in the Accuracy and Coverage Evaluation person universe. The omission of these persons does affect the estimated percentage of person duplication in Census 2000.

Another limitation is that this study concerns duplicates to E-sample persons discovered within the Accuracy and Coverage Evaluation search area. The search area was defined to be the sampled block cluster and one ring of surrounding blocks in clusters chosen for extended search. Within this search area, the duplicate search was performed in housing units and not in group quarters. Person followup found that some E-sample and P-sample persons were Census Day residents of an address outside of the search area. The E-sample person was coded as an 'other residence' erroneous enumeration instead of duplicate and the P-sample person was removed. Research completed after the Accuracy and Coverage Evaluation showed that some E-sample persons coded as correct enumerations and some P-sample persons coded as Census Day residents actually lived outside of the search area.

A related limitation is that this study only makes use of Accuracy and Coverage Evaluation production data. There were subsequent studies of A.C.E. persons that found more person duplicates (see Mule, 2001). The results of these subsequent studies are outside of the scope of this study.

A final limitation is that data from Puerto Rico were not used.

## 4. RESULTS

### 4.1 How did census person duplication in Census 2000 compare with census person duplication in the 1990 Census?

There were some changes from 1990 to 2000 that should be considered when comparing the duplicate percentages :

- In 2000, an operation was developed to reduce the number of duplicate housing units. This operation eliminated some housing unit duplicates before they went to matching. In 1990, there was no such operation. Duplicate housing units contain person duplicates. This means that the person duplicate percentage should be lower in 2000, because some housing unit duplicates were eliminated before 2000 A.C.E. person matching.
- In 2000, the census housing unit duplicate operation reinstated some of the housing units initially flagged as potential duplicates. People enumerated in these reinstated units were excluded from the E sample. There were nearly 2.2 million people who were reinstated, because they were erroneously deleted as duplicates. This means that the A.C.E. did not directly measure the erroneous enumeration rate for nearly 2.2 million people that were in the E-sample universe. In 1990, a smaller number of people were imputed as late census data. The reinstated people were treated the same as a whole person imputation in the dual system estimator.
- There was a change in the search area for duplicate persons. In 1990, rural areas had a search area of two rings of surrounding blocks for duplicates. In 2000, the search area was limited to the block cluster. However, in 2000, there was a targeted extended search that expanded the search area for duplicates to the first ring of surrounding blocks for clusters likely to benefit from an expanded search area. This means that a case could have been considered a duplicate in 1990 and a correct enumeration or a non duplicated erroneous enumeration in 2000.
- There was a change in the assignment of probability of erroneous enumeration when at least one duplicate record was in the E-sample and at least one duplicate record was in a subsampled out housing unit. See the Appendix for information on the calculation of erroneous enumeration probabilities for duplicates.
- The 1990 search area included persons living in non-institutional group quarters while the 2000 search area did not.

**Table 1** gives the aggregate weighted rate of duplication in the E-sample for the 1990 Post Enumeration Survey (PES) and the 2000 A.C.E. Here, rates are of the total weighted number of persons in the E-sample and are expressed in percentages. The percent duplication is lower in 2000, but this is largely attributable to the fact that housing units deleted and reinstated by the duplicate housing unit operation in 2000 are not included in the computation of the 2000 A.C.E. person duplication percentage. These deleted and reinstated units had person duplicates in them so that the decrease in percent duplication would have been smaller had these deleted and reinstated units been included in 2000. Furthermore, the 2000 percentage is based upon A.C.E. production data only. There were subsequent studies of A.C.E. persons that found more person duplicates (see Mule, 2001). However, the percentage given in Table 1 does not include the results of these studies.

**Table 1: E-sample Overall Percent Person Duplication**

<b>Year/Survey</b>	<b>Weighted Percent Duplication</b>	<b>Estimated Number of Duplicates</b>
1990 PES	1.62	4.09 million
2000 ACE	0.76	2.01 million

Note: In 2000, 0.87 percent of the census was reinstated housing units.

**4.2 What was the frequency of census person duplication? Did this frequency vary within important variables? What attributes did person duplicates have in common with those that they duplicated? On which attributes did person duplicate pairs disagree?**

**Tables 2, 3, 4, 6, 7, 10, 12, 13, 13a, 14, 15, and 15a** give weighted percentages of person duplication frequency in the E-sample by important variables. The frequency rate for a given variable level is the percentage of the weighted total of E-sample persons in that level that are coded as duplicate. These tables display variable level names, the percentage duplication frequency (percent), the stratified jackknife standard error (s.e.), the rank of the percentage duplication frequency in descending order (rank), and the ranks of levels with which a significant difference was found (differ). Each pair of levels of each variable was compared by a t-test with a critical value that reflects the Bonferroni criterion. These critical values of t are given below each table. Selected tables have the weighted percentage of the E-sample pertaining to the given level in the final column.

**Tables 5, 8, 9, and 11** give cross classifications of linked duplicate pairs by selected variables. Each pair consists of the person record that is duplicated (called the primary) and the duplicate record. The row of each table gives the variable level of the primary record while the column of each table gives the variable level of the duplicate record.

**Table 2** gives weighted duplication percentages by A.C.E. regional office. It shows that the New York and Boston offices have the highest rates of duplication while the Detroit and Los Angeles offices have the lowest. In fact, the New York office had significantly higher duplication percentages than all other regional offices. The Boston office had significantly higher duplication percentages than three of the remaining ten regional offices (excluding New York). It appears that census duplication was more frequent in the Northeast.

<b>Regional Office</b>	<b>Percent</b>	<b>(s.e.)</b>	<b>Rank</b>	<b>Differ From</b>
Boston	1.07	(0.16)	2	1,10,11,12
New York	2.04	(0.14)	1	all
Philadelphia	0.61	(0.06)	8	1
Detroit	0.44	(0.05)	12	1,2,3
Chicago	0.71	(0.07)	6	1
Kansas City	0.59	(0.08)	9	1
Seattle	0.76	(0.10)	4	1
Charlotte	0.69	(0.06)	7	1
Atlanta	0.72	(0.07)	5	1
Dallas	0.82	(0.08)	3	1,12
Denver	0.51	(0.06)	10	1,2
Los Angeles	0.48	(0.05)	11	1,2

Estimated number of E-sample persons: 264,578,863

Critical value of t: 3.164

**Table 3** gives weighted duplication percentages by size of the metropolitan statistical area (MSA). The possible metropolitan area sizes are large, medium, small, and non-MSA. The level non-MSA is a close approximation to rural, sparsely populated locations. **Table 3** shows that large metropolitan areas and rural areas had higher duplication percentages than small and medium metropolitan areas. The results of Table 3 reinforce the results of Table 2 because the New York and Boston regional offices have cities that are in large MSAs.

**Table 3: Weighted Person Duplication Percentages by Size of Metropolitan Area (E-sample)**

<b>MSA Size</b>	<b>Percent</b>	<b>(s.e.)</b>	<b>Rank</b>	<b>Differ From</b>
Large MSA	0.92	(0.05)	2	3,4
Medium MSA	0.56	(0.04)	4	1,2
Small MSA	0.62	(0.06)	3	1,2
Non-MSA	0.95	(0.06)	1	3,4

Estimated number of E-sample persons: 264,578,863

Critical value of t: 2.386

**Table 4** gives weighted duplicate percentages by type of census return. Generally, people either filled out their census forms themselves or a census enumerator filled them out. Internet responses were counted as mail returns while responses obtained through Telephone Questionnaire Assistance (TQA) were counted as enumerator filled returns. Persons living in mailout/mailback areas could have enumerator filled returns if they did not respond by mail. Results show that enumerator filled returns had higher duplication percentages than mail returns.

**Table 4: Weighted Person Duplication Percentages by Type of Census Return (E-sample)**

<b>Type of Return</b>	<b>Percent</b>	<b>(s.e.)</b>	<b>Rank</b>	<b>Differ</b>
Mail	0.41	(0.02)	2	1
Enumerator	1.89	(0.08)	1	2

Estimated number of E-sample persons: 264,578,863

Critical value of t: 1.65

**Table 5** cross classifies primary-duplicate pairs by type of census return. Again, the primary person is usually matched to a P-sample person or is otherwise confirmed to be a correct enumeration. The next to last row of Table 5 represents pairs in which the duplicated person was enumerated in a surrounding block of a Targeted Extended Search (TES) cluster and the next to final column of Table 5 represents pairs where the duplicate person was enumerated in a housing unit that was subsampled out of the E sample. Results indicate that 56.1 percent of all pairs are the result of one mail and one enumerator return.

<b>Return of Primary</b>	<b>Return of Duplicate</b>			
	Mail	Enumerator	Sub-sampled	Total
Mail	640	2,210	404	3,254
Enumerator	1,648	1,250	489	3,387
Surrounding Block	75	167	0	242
Total	2,363	3,627	893	6,883

**Table 6** gives the weighted duplication percentages by race domain. Non-Hispanic Blacks and Hispanics had the highest duplication percentages while Non-Hispanic Whites had the lowest. African Americans and Hispanics had significantly higher duplication rates than Non-Hispanic Whites and American Indians off reservations.

**Table 6: Weighted Person Duplication Percentages by Racial/Ethnic Domain (E-sample)**

<b>Domain</b>	<b>Percent</b>	<b>(s.e.)</b>	<b>Rank</b>	<b>Differ From</b>
American Indian on reservation	0.74	(0.15)	5	none
American Indian off reservation	0.65	(0.15)	6	1
Hispanic	1.12	(0.07)	2	7
Non-Hispanic Black	1.19	(0.08)	1	6,7
Native Hawaiian or Pacific Islander	0.76	(0.21)	4	none
Non-Hispanic Asian	1.02	(0.17)	3	none
Non-Hispanic White	0.61	(0.03)	7	1,2

Estimated number of E-sample persons: 264,578,863

Critical value of t: 2.815

**Table 7** gives the weighted duplication percentages by age and sex. It shows that the duplication percentage was significantly lower among children (age 0-17). Males aged 18-29 have higher duplication percentages than males of any other age group. In the 30-49 age group, the duplication rate for males is significantly higher than the rate for females. Females aged 18-29 have higher rates than females aged 30-49.

**Table 7: Weighted Person Duplication Percentages by Age/Sex (E-sample)**

<b>Age/sex</b>	<b>Percent</b>	<b>(s.e.)</b>	<b>Rank</b>	<b>Differ From</b>
0-17	0.58	(0.03)	7	all
18-29 Male	1.01	(0.06)	1	4,5,6,7
18-29 Female	0.88	(0.06)	2	6,7
30-49 Male	0.82	(0.04)	4	1,6,7
30-49 Female	0.70	(0.03)	6	1,2,4,7
50+ Male	0.79	(0.04)	5	1,7
50+ Female	0.84	(0.06)	3	7

Estimated number of E-sample persons: 264,578,863

Critical value of t: 2.807

**Table 8** cross classifies the duplicate pairs by age grouping only. Pairs in which both the age of the primary and the age of the duplicate are missing are considered to be not in agreement. It shows that 77.4 percent of all pairs agree on age grouping. **Table 9** cross classifies the duplicate pairs by gender only. Pairs in which both the gender of the primary and the gender of the duplicate are missing are considered to be not in agreement. It shows that 93.3 percent of all pairs agree on gender.

**Table 8: Cross Classification of Duplicate E-sample Pairs by Age Category**

Age of Primary	Age of Duplicate					Total
	0-17	18-29	30-49	50+	Missing	
0-17	1,177	36	10	11	116	1,350
18-29	34	1,090	43	7	131	1,305
30-49	20	49	1,519	62	233	1,883
50+	10	11	67	1,543	306	1,937
Missing	63	66	122	93	64	408
Total	1,304	1,252	1,761	1,716	850	6,883

**Table 9: Cross Classification of Duplicate E-sample Pairs by Gender**

Gender of Primary	Gender of Duplicate			Total
	Male	Female	Missing	
Male	3,227	107	74	3,408
Female	115	3,197	65	3,377
Missing	45	40	13	98
Total	3,387	3,344	152	6,883

**4.3 What kinds of housing units contained person duplicates? To what extent did person duplication reflect housing unit duplication? Were there other census operations contributing to person duplication?**

Next we investigate the housing unit characteristics of census duplicates. **Table 10** gives the weighted person duplication rates by housing tenure of person. It shows that duplication percentages were significantly higher among nonowners. **Table 11** cross classifies the housing tenure of linked duplicate pairs. Pairs in which both the tenure of the primary and the tenure of the duplicate are missing are considered to be not in agreement. It shows that 11.0 percent of all pairs are owner on one and nonowner on the other.

**Table 10: Weighted Person Duplication Percentages by Housing Tenure (E-sample)**

Tenure	Percent	(s.e.)	Rank	Differ From
Nonowner	1.24	(0.06)	1	2
Owner	0.55	(0.02)	2	1

Critical value of t: 1.645

Estimated number of E-sample persons: 264,578,863

**Table 11: Cross classification of Duplicate E-sample Pairs by Housing Tenure**

Tenure of Primary	Tenure of Duplicate			
	Owner	Nonowner	Missing	Total
Owner	2,357	377	290	3,024
Nonowner	379	2,770	269	3,418
Missing	150	179	112	441
Total	2,886	3,326	671	6,883

**Table 12** gives weighted person duplication percentages by the housing unit enumeration status as determined in the final housing unit match. Housing unit enumeration status can be divided into correct enumerations which include matched units, housing units that are duplicates of other housing units, other erroneous enumerations which include geocoding errors, and units with unresolved enumeration status. Results show that person duplication frequency percentages were significantly higher in duplicate housing units and significantly lower in correctly enumerated housing units.

**Table 12: Weighted Person Duplication Percentages by Housing Unit Enumeration Status (E-sample)**

<b>Housing Unit Enumeration Status</b>	<b>Percent</b>	<b>(s.e.)</b>	<b>Rank</b>	<b>Differ From</b>
Correctly Enumerated	0.54	(0.02)	4	all
Duplicate Housing Unit	32.88	(3.48)	1	all
Unresolved	7.89	(2.24)	3	1,4
Other Erroneous Enumerations	10.37	(0.94)	2	1,4

Estimated number of E-sample persons: 264,578,863

Critical value of t: 2.386

**Table 13** gives weighted person duplication percentages by type of structure. It shows that the small multiunits having 2-9 units at basic street address had significantly higher duplication percentages than single family homes and larger apartment buildings. **Table 13a** divides the small multiunits into those having exactly 2 units at basic street address and those having from 3-9 units at basic street address. It shows that addresses having exactly 2 units at basic street address have a significantly higher duplication percentage than addresses with 3-9 units at basic street address.

**Table 13: Weighted Person Duplication Percentages by Type of Structure (E-sample)**

<b>Number of Units at Address</b>	<b>Percent</b>	<b>(s.e.)</b>	<b>Rank</b>	<b>Differ From</b>
1	0.37	(0.02)	3	all
2-9	3.33	(0.15)	1	all
10+	0.86	(0.04)	2	all

Estimated number of E-sample persons: 264,578,863

Critical value of t: 2.121

**Table 13a: Weighted Person Duplication Percentages by Type (E-sample)**

<b>Number of Units at Address</b>	<b>Percent</b>	<b>(s.e.)</b>	<b>Rank</b>	<b>Differ From</b>
2	4.50	(0.32)	1	2
3-9	2.78	(0.20)	2	1

Estimated number of E-sample persons: 264,578,863

Critical value of t: 1.645

**Table 14** gives weighted person duplication percentages by the prevalence of mobile homes in the block cluster. Results show that duplication percentages were low in clusters with few or no mobile homes. There is a jump in the rate of duplication in clusters with a medium to high number of mobile homes. This jump does reflect a significant difference.

**Table 14: Weighted Person Duplication Percentages by Percentage of Mobile Homes in Cluster (E-sample)**

Percent Mobile Homes	Rate	(s.e.)	Rank	Differ From
No mobile homes in cluster	0.69	(0.03)	4	1,2
Less than 10% in cluster	0.74	(0.05)	3	1,2
From 10% to 50% in cluster	1.31	(0.11)	1	3,4
More than 50% in cluster	1.30	(0.21)	2	3,4

Estimated number of E-sample persons: 264,578,863

Critical value of t: 2.234

**Table 15** gives the weighted E-sample person duplicate percentages by source of address. The source of address is the first operation or file in which the address occurred. Addresses are added to the census by a variety of operations that occurred both before and during the census. Major sources of census addresses included:

- **1990 ACF:** These are addresses on file at the Census Bureau in 1990
- **Address Listing (AL):** This was a field operation occurring in non-mailout/mailback enumeration areas
- **Postal Delivery Sequence Files (DSF):** This is a monthly update of addresses from the Postal Service.
- **Block Canvassing (BC):** This was a field verification of addresses on the Master Address File as of January 1999.
- **Local Update of Census Addresses (LUCA):** An update attributable to a cooperative effort with local governments.
- **Questionnaire Delivery (QD):** A field operation where enumerators hand deliver forms to housing units and update addresses in the process.

- **Non Response Followup (NRFU):** These are address updates from enumerators visiting households that have not completed mail returns.
- **Coverage Improvement Followup (CIFU), Be Counted, Telephone Questionnaire Assistance:** These are other Census operations that furnish addresses
- **New Construction (NC):** These are address updates from housing units recently built.
- **Special Place or Group Quarters (SPGQ):** The address is from the census enumeration of special places and group quarters.

Results show that the percentage of duplication is higher in addresses that are from the NRFU and CIFU operation than from other operations occurring in Census 2000, with the exception of New Construction and Special Place or Group Quarters. The final column of Table 15 gives the weighted percentage of E-sample persons living in housing units with the listed source.

**Table 15: Weighted Person Duplication Percentages by Source of Census Address (E-sample)**

Source of Address	Percent	(s.e.)	Rank	Differ From	Percent of E sample
1990 ACF	0.58	(0.03)	10	2,3,5,6,7	63.55
AL	0.72	(0.05)	8	2,3,5,6,7	15.96
DSF	0.61	(0.05)	9	2,3,5,6,7	15.02
BC	2.99	(0.38)	6	2,3,8,9,10	1.99
LUCA	2.32	(0.36)	7	2,3,8,9,10	1.74
QD	3.73	(0.47)	5	8,9,10	1.39
NRFU	8.74	(1.62)	3	6,7,8,9,10	0.19
CIFU	9.07	(1.70)	2	6,7,8,9,10	0.11
NC	5.17	(3.10)	4	none	0.04
SPGQ	9.87	(6.30)	1	none	0.01

Estimated number of E-sample persons: 264,578,863

Critical value of t: 3.051

While Table 15 gives the percentage of Source of Census Address that were person duplicates, **Table 15a** gives the percentage of Source of Census Address that were primaries to person duplicates. The 1990 ACF, Address Listing, Delivery Sequence File, and New Construction had the lowest percentages while NRFU, CIFU, and SPGQ had the highest.

**Table 15a: Weighted Percentages of Source of Census Address Persons that were Primaries to E-sample Person Duplicates**

Source of Address	Percent Primaries*	(s.e.)	Rank	Differ From	Percent of E sample
1990 ACF	0.55	(0.02)	9	3,4,5,6,7,10	63.55
AL	0.80	(0.06)	7	4,8,9,10	15.96
DSF	0.57	(0.05)	8	3,4,5,6,7,10	15.02
BC	1.39	(0.19)	6	8,9,10	1.99
LUCA	1.47	(0.26)	5	8,9,10	1.74
QD	1.73	(0.27)	4	7,8,9,10	1.39
NRFU	3.88	(0.96)	3	7,8,9,10	0.19
CIFU	3.89	(1.19)	2	10	0.11
NC	0.00	(0.00)	10	all but 1	0.04
SPGQ	4.87	(3.76)	1	none	0.01

Estimated number of E-sample persons: 264,578,863

Critical value of t: 3.051

\*This percent does not include primaries with duplicate links to Non E-sample persons.

**4.4 What was the frequency of P-sample person duplication? Did this frequency vary within important P-sample person variables?**

The overall weighted percentage of P-sample persons that were coded as duplicate was 0.22 percent. This is lower than the corresponding figure for E-sample persons (0.76 percent).

**Table 16** gives weighted P-sample person duplication percentages by region. The P-sample person duplication percentages were the lowest in the Midwest.

**Table 16: Weighted P-sample Person Duplication Percentages by Region**

<b>Region</b>	<b>Percent</b>	<b>(s.e.)</b>	<b>Rank</b>	<b>Differ From</b>
Northeast	0.22	(0.02)	3	4
Midwest	0.14	(0.01)	4	1,2,3
South	0.26	(0.02)	1	4
West	0.25	(0.02)	2	4

Critical value of t: 2.386

**Table 17** gives P-sample person duplication percentages by type of enumeration area (TEA). Results show that mailout/mailback (0.22 percent) and update leave (0.21 percent) areas had significantly higher person duplication percentages than urban update leave (0.03 percent) areas.

**Table 17: Weighted P-sample Person Duplication Percentages by Type of Enumeration Area**

<b>TEA</b>	<b>Percent</b>	<b>(s.e.)</b>	<b>Rank</b>	<b>Differ From</b>
Mailout Mailback	0.22	(0.01)	3	6
Update/Leave	0.21	(0.02)	4	6
List/Enumerate	0.16	(0.13)	5	none
Rural Update/Enumerate	0.58	(0.23)	2	none
Urban Update/Leave	0.03	(0.03)	6	3,4
Urban Update Enumerate	1.00	(1.11)	1	none
Mailout Mailback to UL	0.16	(0.12)	5	none

Critical value of t: 2.815

**Table 18** gives person duplication percentages by type of basic address. It shows that the person duplication percentage in mobile homes located outside trailer parks (0.45 percent) is the highest, and it is significantly higher than that within single family homes (0.19 percent) and multiunits in special places (0.06 percent).

**Table 18: Weighted P-sample Person Duplication Percentages by Type of Basic Address**

Type of Basic Address	Percent	(s.e.)	Rank	Differ From
Single family house	0.19	(0.01)	5	1,3
Address with 2 or more units	0.32	(0.03)	3	5,6
Mobile Home not in Park	0.45	(0.08)	1	5,6
Mobile Home in Park	0.22	(0.06)	4	none
Single family unit in Special Place	0.33	(0.23)	2	none
Multiunit in Special Place	0.06	(0.07)	6	1,3

Critical value of t: 2.815

**Table 19** gives P-sample person duplication percentages by gender. There is no significant difference in duplication percentage between males and females.

**Table 19: Weighted P-sample Duplication Percentages by Gender**

Gender	Percent	(s.e.)	Rank	Differ From
Male	0.23	(0.01)	1	none
Female	0.21	(0.01)	2	none

Critical value of t: 1.65

**Table 20** gives P-sample person duplication percentages by age. The duplication percentage of the group aged 18-29 (0.34 percent) was significantly higher than that of the other age groups.

**Table 20: Weighted P-sample Person Duplication Percentages by Age**

Age	Percent	(s.e.)	Rank	Differ From
0-17	0.24	(0.02)	2	1,4
18-29	0.34	(0.02)	1	2,3,4
30-49	0.20	(0.01)	3	1
50+	0.16	(0.01)	4	1,2

Critical value of t: 2.386

**Table 21** gives P-sample person duplication percentages by housing tenure. The duplication percentage of owners was significantly lower than that of renters. This result is analogous to the result for the E sample.

**Table 21: Weighted P-sample Duplication Percentages by Housing Tenure**

Tenure	Percent	(s.e.)	Rank	Differ From
Owner	0.21	(0.01)	2	1
Nonowner	0.26	(0.02)	1	2

Critical value of t: 1.65

**Table 22** gives P-sample person duplication percentages by proxy status. The duplication percentage of persons whose characteristics were reported by proxy was significantly higher than that of those who self report their characteristics.

**Table 22: Weighted P-sample Duplication Percentages by Proxy Status**

Status	Percent	(s.e.)	Rank	Differ From
Proxy	0.37	(0.04)	1	2
Nonproxy	0.21	(0.01)	2	1

Critical value of t: 1.65

## 5. CONCLUSIONS AND RECOMMENDATIONS

The objective of this study was to document the extent of census and P-sample person duplication, to identify the characteristics of persons most likely to be duplicates, and to identify how persons with duplicate links compare to one another. Results of this study can be used to identify characteristics and geographic areas that may be most beneficial to study or target when searching for person duplicates. The results can be used to guide unduplication efforts.

The major conclusions are as follows:

**Census person duplication occurred more frequently than P-sample person duplication.** In Census 2000, 0.76 percent of all E-sample persons were duplicates in while only 0.22 percent of all P-sample persons were duplicates. These figures may reflect the greater complexity of the Census, compared to the Accuracy and Coverage Evaluation. P-sample person information was collected from a single source, the person interview. Census person information was obtained through a greater variety of sources. These results show that it is beneficial to conduct duplicate person searches, and that successful efforts to unduplicate persons can result in better population and coverage estimates.

**Person duplication was not uniform.** Duplication percentages varied, sometimes rather widely, by many variables including race, age/sex category, size of metropolitan area, and type of census return. This shows that it is potentially beneficial to target certain subgroups and areas in conducting person duplicate searches.

The next major conclusions relate to census person duplicates:

**Census person duplication does not only occur in duplicate housing units.** Table 12 indicates that 32.88 percent of the people living in duplicate housing units were coded as person duplicates. This suggests that many person duplicates can be found in duplicate housing units. However, the table also shows that person duplicates occur in other erroneously enumerated housing units, and units with unresolved enumeration status, and in correctly enumerated housing units. **This suggests that some person duplication will be removed when housing units are unduplicated. However, a separate effort to unduplicate persons should complement efforts to unduplicate housing units to achieve the biggest reduction in person duplication.**

**Census duplication is more prevalent in small multiunit housing structures. It is less prevalent in single family homes. It also appears to be more prevalent in mobile homes.** Table 13 shows that there was a significantly higher percentage of person duplicates in units with 2-9 housing units at basic street address (3.33 percent) than in single family housing units (0.37 percent). Table 13a shows that within small multiunits addresses with 2 housing units have significantly higher duplication percentages (4.50 percent) than addresses with 3-9 housing units (2.78 percent). Table 14 indicates that there was a significantly higher percentage of person

duplicates in clusters with more than ten percent mobile homes (1.3 percent) than in clusters with less than ten percent mobile homes (between 0.69 and 0.74 percent). **Future unduplication efforts should be focused on small multiunits and mobile homes.**

**Racial and ethnic groups that are traditionally undercounted are also more frequently duplicated.** African Americans and Hispanics have the highest duplication percentages. Because person duplicates are erroneous enumerations, they actually reduce the census undercount.

## 6. REFERENCES

- Barrett, D, Beaghen, M., Smith, Damon, Burcham, J. (2003). “*Final Draft Report for the Census 2000 Housing Unit Coverage Study, O.3*”, November 5, 2002.
- Childers, D. (2001). “*The Design of the Census 2000 Accuracy and Coverage Evaluation (A.C.E.)*” DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-01, U.S. Census Bureau, January 26, 2001.
- Fay, R. (1990). “VPLX: Variance Estimates for Complex Samples,” *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp 266-271.
- Fay, R. (2001). “*ESCAP II: Evidence of Additional Erroneous Enumeration from the Person Duplication Study*”, “*Executive Steering Committee for A.C.E. Policy (ESCAP II), Report 9*”, U.S. Census Bureau, October 26, 2001.
- Feldpausch, R. (2001). “*ESCAP II: E-Sample Erroneous Enumerations*”, “*Executive Steering Committee for A.C.E. Policy (ESCAP II), Report 5*”, U.S. Census Bureau, October 14, 2000.
- Hocking, R.R. (1986). *Methods and Applications of Linear Models: Regression and the Analysis of Variance* (New York: John Wiley and sons), pp 108-9.
- Hogan, H. (1993). “The 1990 Post-Enumeration Survey: Operations and Results,” *Journal of the American Statistical Association*, Vol 88, No 423, pp 1047-1060.
- Mule, Tom (2001). “*ESCAP II: Person Duplication in Census 2000*”, “*Executive Steering Committee for A.C.E. Policy (ESCAP II), Report 20*”, U.S. Census Bureau, October 11, 2000.
- Nash, Fay (2000). “*Overview of the Duplicate Housing Unit Operation*”, internal document, U.S. Census Bureau, November 7, 2000.
- Raglin, David (2001). “*ESCAP II: Effect of Excluding Reinstated Census People from the A.C.E. Person Process*”, “*Executive Steering Committee for A.C.E. Policy (ESCAP II), Report 13*”, U.S. Census Bureau, October 9, 2000.

## APPENDIX

### Calculating erroneous enumeration rates in 1990 and 2000

Census duplicates are erroneous enumerations and therefore contribute to the rate of erroneous enumeration. The erroneous enumeration rate is the complement of the correct enumeration rate used in forming dual system population estimates. These duplicates are linked with those persons that they duplicate. It is possible for a census person to have more than one duplicate. When two or more census person records refer to the same person, probabilities of erroneous enumeration must be assigned to each record so that the sum of these probabilities is one or less. These probabilities are then used in the calculation of the erroneous enumeration rate. There are differences between the 1990 Post Enumeration Survey and the 2000 Accuracy and Coverage Evaluation in the way in which such probabilities were assigned.

Both the 1990 PES and the 2000 A.C.E. consisted of samples of census block clusters. In both surveys, housing units in sampled block clusters were subsampled. Persons enumerated in housing units in sample after subsampling are said to be in the E sample (E-sample indicator 1), while persons enumerated in housing units out of sample after subsampling are said to have E-sample indicator 2. In both surveys, a duplicate search of E-sample persons was performed both in the E-sample and in the housing units that were out of sample after subsampling. The person duplicates could be in the E-sample, or be enumerated in housing units that were out of sample after subsampling. Each survey assigned probabilities of erroneous enumeration to all records referring to the same person in the following way:

**When all records were in the E-sample:** Each survey assigned a probability of erroneous enumeration of 1 to the duplicate records and a probability of erroneous enumeration of 0 to the record that was duplicated.

**When all duplicate records were in subsampled out housing units:** In each survey the duplicated person was in the E sample. This person was assigned a probability of erroneous enumeration of  $d/d+1$ , where  $d$  was the number of duplicate records. For example, a person with 2 duplicates received a probability of erroneous enumeration equal to  $2/3$ . The duplicate records were not used in computing the enumeration rate, since they were out of the E sample.

**When at least one duplicate record was in the E-sample and at least one duplicate record was in a subsampled out housing unit:** Again, in each survey the duplicated person was in the E sample. The strategies in 1990 and 2000 are listed below:

- **1990:** The duplicated person was assigned a probability of erroneous enumeration of  $d/d+1$ , where  $d$  is the number of duplicates in subsampled out housing units. For example, an E-sample person with one E-sample duplicate and one duplicate in a subsampled out housing unit would receive a probability of erroneous enumeration of  $1/2$

- **2000:** The duplicated person was assigned a probability of erroneous enumeration of  $d/d+e+1$ , where  $d$  is the number of duplicates in subsampled out housing units, and  $e$  is the number of E-sample duplicates. For example, an E-sample person with one E-sample duplicate and one duplicate in a subsampled out housing unit would receive a probability of erroneous enumeration equal to  $1/3$ . There were 11 of these cases in 2000.