CENSUS 2000 E-SAMPLE ERRONEOUS ENUMERATIONS

Roxanne Feldpausch[1]
Bureau of the Census, Washington, DC 20233

Key words: Census 2000, E-sample, Erroneous Enumerations

1.0 Introduction

The goal of Census 2000 was to count everyone in the U.S. in their proper household. However, this did not always happen. To assess the coverage of the census, the Census Bureau undertook the Accuracy and Coverage Evaluation (A.C.E.). The A.C.E. determined whether people in the E-sample, a sample of the people counted by the census in housing units, were correctly enumerated or erroneously enumerated.

To determine the number of erroneous enumerations, the E-sample people were matched to the people captured in the A.C.E. Computer and clerical matching classified E-sample people as matched, not matched, or possibly matched. A person who was matched was captured in both the census and A.C.E. The nonmatched and possibly matched people were followed-up to determine if they were correctly or erroneously enumerated in the block cluster (a group of geographically contiguous blocks) according to census residence rules. That is, the people who were correctly enumerated were people who the census correctly captured in the block cluster. Erroneously enumerated people were people that the census captured in error in the block cluster. If the follow-up interview could not determine the person to be correctly or erroneously enumerated, the enumeration status for the E-sample person was unresolved. Those people with unresolved enumeration status had their probability of correct enumeration imputed based on those cases that were successfully followed-up. Those who matched were considered correctly enumerated. See Childers (January, 2001) for more details.

The rate of erroneous enumerations for a given post-stratum is related to the dual system estimates, the estimate of the population count using census and A.C.E. data. Assuming everything else is held constant, as the erroneous enumeration rate increases the dual system estimate decreases. Dual system estimates allow us to calculate undercounts, which is an important measure of the quality of the census. Understanding erroneous enumerations will help us understand the quality of the census. Knowing which variables are related to a person being erroneously enumerated will also aid in the planning for the 2010 Census.

Section 2, discusses the various types of erroneous enumerations and gives their definitions. Section 3 gives the methodology used in this analysis. Section 4 shows the analysis of erroneous enumeration rates by related variables. Section 5 summarizes the results.

2.0 Definitions

The E-sample consists of data-defined people in selected housing units. To be data-defined, a person record has to have at least two characteristics, where name counts as a characteristic. According to A.C.E. rules and definitions, there were five types of erroneous enumerations:

- duplicates
- other residence
- fictitious
- insufficient information for matching
- geocoding errors

Duplicates: The census counted the same person more than once. Duplicates could happen on the same form, on a different form at the same address, at a different address in the same block cluster or at a different address in a surrounding block.

Other residence: The A.C.E. person follow-up interview determined that the E-sample person was not a resident on census day because the person should have been enumerated at an other residence. This includes people who were duplicate outside of the A.C.E. search area.

Fictitious: The E-sample nonmatch was determined to be fictitious in this cluster during the A.C.E. person

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

follow-up interview. The person may have existed elsewhere, but the interviewer could not find anyone in the cluster who knew the person. The interviewer had to talk to at least three knowledgeable people in the cluster before a person could be considered fictitious.

Insufficient information for matching and follow-up: To have sufficient information for matching and follow-up, an E-sample person had to have a complete name and at least two other characteristics. People with insufficient information for matching and follow-up were people whose name was blank or invalid, or people who had only one other characteristic.

Geocoding errors: If the census housing unit existed outside the A.C.E. search area, all of the people in the housing unit were erroneous enumerations due to geocoding error.

There are also people with either unresolved match or residence status. We did not get enough information in A.C.E. person follow-up to determine either their match or residence status for the E-sample person. These people have their probability of correct enumeration imputed. In sections 4.1, 4.2, and 4.3 I combined all of the people with unresolved status into a category called unresolved. In section 4.4, I put the people with unresolved status into categories based on how they were imputed.

It should be noted that a different definition of what is considered erroneous would lead to a different erroneous enumeration rate.

3.0 Methods

The erroneous enumeration rate is the weighted number of people in the E-sample that were erroneously enumerated divided by the total weighted number of people in the E-sample. The probability of erroneous enumeration (one minus the probability of correct enumeration) was used to determine the number of erroneously enumerated people, I used. Rates for the different types of erroneous enumerations were calculated similarly, with the numerator being the number of that type of erroneous enumerations and the denominator being the total number of people in the E-sample.

Stratified Jackknife method and VPLX were used to compute the standard errors. All hypothesis testing were two-tailed at a 0.10 significance level. Bonferroni's adjustment was used for multiple comparisons. This is an analysis of the United States, it includes the 50 states and the District of Columbia.

4.0 Results

Section 4.1 compares Census 2000 results with the 1990 Census. Erroneous enumeration rates by the post-stratification variables are examined in Section 4.2. Section 4.3 gives erroneous enumeration rates by other interesting variables. The various types of erroneous enumerations are discussed in Section 4.4.

4.1 Comparison of Census 2000 data with 1990

In 1990 Post-Enumeration Survey (PES) there were similar types of erroneous enumeration categories to those measured by 2000 A.C.E. However, there were differences between the PES and the A.C.E. that limits our ability to make comparisons. For example, in Census 2000, there was a Housing Unit Duplication Operation that involved the removal of duplicate housing units and people. This helps explain why the duplicate rates appear to have decreased from 1990 to 2000. See Feldpausch (2001) for a more complete list of the differences.

Table 1 gives a comparison of 2000 results with 1990. The distribution of erroneous enumerations in 2000 looks different than in 1990. The percent other residence appears lower in 2000 than in 1990. We are still looking into possible explanations for this difference.

Table 1 Comparison of Type of Erroneous Enumerations

	2000	**1990**
Duplicate	0.8	1.6
Other Residence	1.0	2.2
Fictitious	0.3	0.2
Insufficient Info.	1.8	1.2
Geocoding	0.2	0.3
Unresolved	0.6	0.3
Total	**4.7**	**5.8**

Note: 1990 data are from Childers (September, 2001) and related to the PES universe

4.2 Erroneous Enumeration Rate by Post-stratification Variables

This section describes the types of erroneous enumerations for the following post-stratification variables: return rate, race/ethnicity, age/sex, tenure, place size and type of enumeration, and region.

Return Rate

Return rates were an important indicator of public cooperation with the census. Tract-level return rates were calculated for each tract with mailback enumeration areas. Areas with high return rates were expected to have lower rates of erroneous enumerations than areas with low return rates. Return rate was an A.C.E. post-stratification variable for the Non-Hispanic White or "Some other race," Non-Hispanic Black, and Hispanic domains. Therefore, E-sample persons in these three race/Hispanic origin domains were affiliated with a high or low return rate indicator value. E-sample persons in all other race/Hispanic origin domains were assigned a return rate indicator value of "Not Applicable" since they were not post-stratified by return rate. See Haines (2001) for details on return rate calculations and the high/low designation.

Table 2 shows that E-sample persons associated with high return rate indicator values had a lower erroneous enumeration rate than both E-sample people with low return rate indicator values and E-sample persons who were not post-stratified by the return rate variable.

Table 2 Percent Erroneous Enumerations by Return Rate (standard errors)	
High	4.20 (0.08)
Low	6.15 (0.14)
Not Applicable	5.52 (0.27)
Total	**4.72 (0.07)**

Race/Ethnicity

For post-stratification purposes, there are seven race/ethnicity groups: American Indian on reservation, American Indian off reservation, Hispanic, Non-Hispanic black, Native Hawaiian or Pacific Islander, Non-Hispanic Asian and Non-Hispanic white. Haines (2001) explains how multi-racial people were placed into categories. See Table 3 for the percent of erroneous enumerations.

The following is a brief explanation of which race/ethnicity groups were significantly different with respect to total erroneous enumeration rate:

- **American Indians on reservations** had a lower rate than American Indians off reservations, non-Hispanic blacks and Hispanics.
- **American Indians off reservations** had a higher rate than American Indians on reservations and non-Hispanic whites.
- **Hispanics** had a higher rate than American Indians on reservations and non-Hispanic whites. They had a lower rate than non-Hispanic blacks.
- **Non-Hispanic blacks** had a higher rate than American Indians on reservations, non-Hispanic Asians, non-Hispanic whites and Hispanics.
- **Hawaiians and Pacific Islanders** had a higher rate than non-Hispanic whites.
- **Non-Hispanic Asians** had a higher rate than non-Hispanic whites and had a lower rate than non-Hispanic blacks.
- **Non-Hispanic Whites** had a lower rate than all other categories except American Indians on reservations.

Table 3 Percent Erroneous Enumerations by Race/Ethnicity (standard errors)	
1. American Indian on reservation	4.19 (0.34)
2. American Indian off reservation	6.03 (0.56)
3. Hispanic	5.54 (0.18)
4. Non-Hispanic black	7.27 (0.21)
5. Hawaiian or Pacific Islander	6.95 (1.00)
6. Non-Hispanic Asian	5.43 (0.32)
7. Non-Hispanic white	4.10 (0.02)
Total	**4.72 (0.07)**

Age/Sex

In the past, people 18-29 years of age were difficult to count. One reason is that they tended to be more mobile than other age categories. See Table 4 for the percent of erroneous enumerations broken down by the age/sex categories used for post-stratification. The following explains which values were significantly different:

- **0-17 years of age** had a lower rate than all other categories except 30-49 females.
- **18-29 males** had a higher rate than all other categories.
- **18-29 females** had a higher rate than 0-17 years of age, 30-49 males, 30-49 females, 50+ males and 50+ females. They had a lower rate than 18-29 males.
- **30-49 males** had a higher rate than 0-17 years of age and 30-49 females. They had a lower rate than 18-29 males and 18-29 females.
- **30-49 females** had a lower rate than all other categories except 0-17 year olds.
- **50+ males** had a higher rate than 0-17 years of age and 30-49 females. They had a lower rate than 18-29 males and 18-29 females.

- **50+ females** had a higher rate than 0-17 years of age and 30-49 females. They had a lower rate than 18-29 males and 18-29 females.

The breakdown of age and sex shown here is based on the breakdown for post-stratification purposes. There was some concern about whether the 0-17 males and females differed with respect to their probability of erroneous enumeration. It is interesting to note that 0-17 males were not significantly different from 0-17 females in total erroneous enumeration rate or by any of the different types of erroneous enumerations.

Table 4 Percent Erroneous Enumerations by Age/Sex (standard errors)	
0-17 Male and Female	4.06 (0.09)
18-29 Male	7.13 (0.16)
18-29 Female	6.39 (0.15)
30-49 Male	4.77 (0.11)
30-49 Female	3.99 (0.09)
50+ Male	4.66 (0.11)
50+ Female	4.49 (0.10)
Total	**4.72 (0.07)**

Tenure (Owner vs Non-Owner)

We expected that owners would have a lower erroneous enumeration rate than non-owners. Owners tend to live in the same place longer and have more connections to the community than non-owners. In Census 2000, owners had a significantly lower erroneous enumeration rate than non-owners (see Table 5). They had a significantly lower duplicate rate, fictitious rate, insufficient information rate, other residence rate and unresolved rate than non-owners. The only rate in which owners and non-owners did not differ was the geocoding error rate.

Table 5 Percent Erroneous Enumerations by Tenure (standard errors)	
Owner	3.59 (0.08)
Non-Owner	7.31 (0.13)
Total	**4.72 (0.07)**

Place Size and Type of Enumeration

Metropolitan Statistical Areas (MSA) were broken down into four categories: large, medium, small and non-MSA. These MSA categories were combined with information about how people got their form: mailout/mailback (MO/MB) and not mailout/mailback. Table 6 shows that large MSA, mailout/mailback areas had a significantly higher rate of erroneous enumeration than all of the other categories. Large MSA mailout/mailback tended to have a higher duplicate rate, fictitious rate and insufficient information rate than the other categories.

Table 6 Percent Erroneous Enumerations by Place Size and Type of Enumeration (standard errors)	
Large MSA MO/MB	5.22 (0.16)
Medium MSA MO/MB	4.51 (0.12)
Small MSA & Non-MSA MO/MB	4.39 (0.15)
Not MO/MB	4.62 (0.04)
Total	**4.72 (0.07)**

Region

The Census Bureau divided the country into four regions: Northeast, Midwest, South and West. The Midwest had a significantly lower rate of erroneous enumerations than the other regions (see Table 7). It tended to have lower rates of insufficient information, other residence and unresolved than the other regions. The Northeast's duplicate rate was significantly higher than all other regions.

Table 7 Percent Erroneous Enumerations by Region (standard errors)	
Northeast	5.05 (0.16)
Midwest	3.82 (0.13)
South	5.07 (0.14)
West	4.80 (0.15)
Total	**4.72 (0.07)**

4.3 Erroneous Enumeration Rates by Other Variables

Imputation

Some people did not answer all of the census questions. When this happened, we imputed the missing characteristics for the person in the census. There were also cases where the data was edited through consistency edits. People with some imputations or some data edits had a significantly higher erroneous enumeration rate than those people with no imputations and no data edits (see Table 9). People with some imputations or some

data edits had a significantly higher duplicate rate, fictitious rate, insufficient information rate, other residence rate and unresolved rate. In the E-sample, 13.0 percent of the people had some imputation or some data edits.

Table 9 Percent Erroneous Enumerations by Imputation (standard errors)	
No Imputation and No Data Edits	3.26 (0.07)
Some Imputation or Data Edits	14.47 (0.25)
Total	**4.72 (0.07)**

Form Length

Census 2000 had two different lengths of questionnaires: short and long. The short form asked: name, age, sex, race, Hispanic origin and tenure. The long form, filled out by 16.5 percent of the E-sample, asked those questions along with demographic and economic questions. There was no difference in the overall erroneous enumeration rate for people captured using the different form lengths (see Table 10). However, people captured on the long form had a significantly lower other residence rate and unresolved rate than people captured on the short form. The duplicate rate for people captured on the short form was significantly higher than that of the people captured on the long form.

Table 10 Percent Erroneous Enumerations by Form Length (standard errors)	
Short	4.75 (0.08)
Long	4.57 (0.12)
Total	**4.72 (0.07)**

Response Method

Most households were self-reporting (mail returns and internet returns). However, 23.7 percent of people in the E-sample were captured on enumerator filled returns. An enumerator visited housing units in areas without reliable mail delivery, areas with a high percentage of people who used post-office boxes and people who did not mail back their census form. The enumerators tried to get an interview with a household member. Sometimes this was not possible, so the enumerator had to get a proxy interview with someone outside the household. Of the enumerator filled returns, 11.3 percent were with a proxy respondent.

The self-reporting people had a significantly lower erroneous enumeration rate than both the proxy and non-proxy enumerator filled returns (see Table 11). The non-proxy enumerator filled returns had a significantly lower erroneous enumeration rate than the proxy enumerator filled returns. The same pattern held for duplicate rates, insufficient information rates, other residence rates and unresolved rates. The self-reporting people had a significantly lower fictitious rate than both the proxy and non-proxy enumerator filled returns, however proxy and non-proxy enumerator filled returns did not differ from each other.

Table 11 Percent Erroneous Enumerations by Response Method (standard errors)	
Self-reporting	2.90 (0.06)
Enumerator	10.57 (0.19)
Non-proxy	7.09 (0.16)
Proxy	37.22 (0.81)
Total	**4.72 (0.07)**

4.4 The Different Types of Erroneous Enumerations

The breakdown of types of erroneous enumerations can be seen in Table 1. For complete analysis of the types of erroneous enumerations, see Feldpausch (2001). Some highlights are given below:

- Duplicate rates are high in large cities in the Northeast, 1.5 percent.
- Conflicting households, households where the census captured one family and the A.C.E. captured another family, had high fictitious rates, 8.8 percent.
- In all variables analyzed, erroneous enumerations due to geocoding error were insignificant in all tests. Targeted Extended Search procedures reduced the effects of erroneous enumerations due to geocoding error by allowing correct enumerations in the surrounding blocks
- Insufficient information was the highest in enumerator filled returns (4.7 percent), especially if the respondent was a proxy (27.4 percent). American Indians on reservations had low rates of insufficient information, 0.9 percent.
- The Midwest had low rates of other residences, 0.8 percent, and American Indians on reservations had high rates of other residences, 1.7 percent.

- The Boston Regional Office had a very low unresolved rate, 0.2 percent. People 18-29 years of age had a high unresolved rate, 1.3 percent.

People with unresolved status are people for whom we do not get enough information in A.C.E. person follow-up to determine their enumeration status. However, we can gain some information about the enumeration status of these people by looking at their follow-up forms. Based on this information, we can classify them into the type of erroneous enumeration categories. This information was also used to imputed the person's probability of correct enumeration during the missing data procedures (Cantwell, 2001).

Table 12 shows how the rates of the various types of erroneous enumerations change, in 1990 and 2000, when the people with unresolved status are incorporated into the different rates. A Comparison of Table 12 and Table 1 shows that the difference percent other residence has decreased.

Table 12 Reclassification of People with Unresolved Status

	2000	**1990**
Duplicate	0.8	1.7
Fictitious	0.5	0.2
Geocoding Error	0.2	0.4
Other Residence	1.4	2.3
Insufficient Information	1.8	1.2
Total	**4.7**	**5.8**

Note: 1990 data are from Childers (September, 2001) and related to the PES universe

5.0 Summary

The rate of erroneous enumerations decreased from 5.8 percent in 1990 to 4.7 percent in 2000. Much of the difference is due to changes in procedures.

All of the post-stratification variables had categories that differed significantly in their probability of being erroneously enumerated.

To control the number of erroneous enumerations in the future, we should try to limit the number of proxy interviews.

6.0 References

Cantwell, P and Childers, D. (March 2001), "Accuracy and Coverage Evaluation Survey: A Change to the Imputation Cells to Address Unresolved Resident and Enumeration Status**",** DSSD Census 2000 Procedures and Operations Memorandum Series Q-44.

Childers, D. (January 2001), "Accuracy and Coverage Evaluation: The Design Document", DSSD Census 2000 Procedures and Operations Memorandum Series S-DT-01.

Childers, D. (September, 2001), "1990 E-Sample Documentation", Census Bureau Internal Memorandum.

Feldpausch, R. (October, 2001), "Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report Number 5: E-Sample Erroneous Enumeration Analysis", DSSD Census 2000 Procedures And Operations Memorandum Series T-11.

Haines, D. (March, 2001), **"**Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Dual System Estimation (U.S.) -Re-issue of Q-37", DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter Q-48.