ERRONEOUSLY ENUMERATED PEOPLE IN THE CENSUS 2000 DRESS REHEARSAL

Roxanne Feldpausch and Danny R. Childers[1]
Roxanne Feldpausch, Bureau of the Census, Washington, DC 20233

Key words: Integrated Coverage Measurement, Dual System Estimation

I. Introduction

The Census 2000 Dress Rehearsal was conducted in three sites: Sacramento, California; Menominee county, Wisconsin (mainly Menominee Indian Reservation); and Columbia and its surrounding areas in South Carolina. In Sacramento and Menominee, the census was conducted in two parts. First, the initial phase enumerated the people. Then the Integrated Coverage Measurement (ICM) estimated the people missed in the initial phase. The combination was the census count. In South Carolina a traditional census was conducted. After the census, there was a Post-Enumeration Survey to estimate the coverage of the census. In this paper both the initial phase and the traditional census are referred to as a census and both of the coverage surveys are referred to as the ICM.

The ICM estimated, among other things, how many people enumerated in the census were enumerated in error. The number of erroneous enumerations is one of the inputs into the dual system estimator, which is a factor used to determine the final census count (Schindler, 1999). In this paper, we look at various factors which may be related to a person's probability of being erroneously enumerated.

To determine the number of erroneous enumerations in the census, the E-sample people (the people captured in the census) were matched to the people captured in the ICM. After the computer and clerical matching phase, the E-sample people were classified as matched, not matched, or possibly matched. Those who matched were considered correctly enumerated. The nonmatched E-sample people were followed up to determine if they were correctly or erroneously enumerated in the block cluster according to census residence rules. 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 erroneous enumeration probabilities imputed.

Section II discusses the methods used to analyze the data. Section III examines the erroneous enumeration rates of various subgroups. Sections IV examines various types of erroneous enumerations. Section V summarizes the findings.

II. Methodology

For analysis purposes we broke the South Carolina site up into three distinct areas: the city of Columbia, referred to as Columbia; the remaining counties which were mail-out/mail-back, referred to as Other SC; and the counties that were update/leave, referred to as Rural SC. For mail-out/mail-back areas, the mailing list was obtained from the US Postal Service, 1990 Census, local, tribal, and other potential supplementary address sources. A census questionnaire was mailed to the addresses and if occupied, the residents were to mail back a completed form. For update/leave areas the enumeration procedures were different. The address list was obtained by Census Bureau employees who created a listing of addresses before the census. The enumerators updated the census address list and left a questionnaire for the household to complete and mail back to the Census Bureau.

In these five different areas, we examined the erroneous enumeration rates of people in various subsets of the population. We estimated the erroneous enumeration rate by the number of erroneous enumerations divided by the total number of people in the E sample. For the erroneous enumeration rates, we calculated the standard errors using the simple Jackknife method. The simple Jackknife should yield standard errors similar to those produced with the stratified Jackknife which was used in the Dress Rehearsal. These standard errors were computed using the statistical package VPLX. The internet site www.census.gov/sdms/www/vwelcome.html has more information on VPLX.

Once we computed standard errors, we used a two-

1 Roxanne Feldpausch and Danny Childers are mathematical statisticians in the Decennial Statistical Studies Division of the US Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. It is released to inform interested parties of research and to encourage discussion.

tailed t-test to determine which differences were significant. When there were multiple comparisons, we used the Bonferroni adjustment to determine which comparisons were significant (Games, 1971). All tests were conducted at a 0.10 significance level. The tests were conducted under the assumption that the observations were independent of each other. The analyses were conducted on the final data, after weighting and imputation. All numbers in this paper are weighted and the probability of erroneous enumerations for people with unresolved enumeration status are included in the percent of erroneous enumerations.

III. Percentage Erroneous Enumeration

For the 1998 Dress Rehearsal, the rate of erroneous enumerations varied across the areas from 7.7 percent in Columbia to 16.5 percent in Other SC. In the 1990 Census, the erroneous enumeration rate for the nation was 4.6 percent (Griffin and Moriarity, 1992). Table 1 shows the percentage of erroneous enumerations out of the total number E sample people for each site. It also gives the number of erroneous enumerations in each site. To compare the percent of erroneous enumerations among the different areas, we used the Bonferroni adjustment. The adjusted alpha is 0.010 and the criterion t-value is 2.560. Other SC had a significantly higher erroneous enumeration rate than Sacramento (t=3.063), Columbia (t=4.202) and Menominee (t=2.762). The remaining areas did not differ in their erroneous enumeration rates.

Table 1: Percentage (Standard Error) and Number of Erroneous Enumerations by Site

Site	Percent EE	Number of EE
Sacramento	10.5 (0.71)	38,878
Columbia	7.7 (1.02)	6,438
Rural SC	10.3 (1.84)	14,858
Other SC	16.5 (1.82)	58,837
Menominee	9.8 (1.62)	321

One group of characteristics that we examined was the poststrata variables for the dual system estimator. These variables are: tenure, sex, age and race. Tenure has been shown to be related to erroneous enumeration rates. In the 1990 Census, renters were more likely to be enumerated in error than owners (Griffin and Moriarity, 1992). In Sacramento (t=4.818), Rural SC (t=1.691), and Other SC (t=2.386) this held true. For these three areas, owners had significantly lower erroneous enumeration rates than renters. In Columbia (t=0.384) and Menominee (t=1.240) there were no differences between the erroneous enumeration rates of renter and owners. See Table 2 for the percentage of erroneous enumerations by renter and owner.

Table 2: Percentage of Erroneous Enumerations (Standard Error) by Tenure

Site	Owner	Renter
Sacramento	7.8 (0.56)	13.6 (1.20)
Columbia	7.2 (0.63)	8.1 (2.21)
Rural SC	9.2 (1.50)	15.8 (4.74)
Other SC	13.2 (1.24)	24.6 (4.82)
Menominee	11.5 (1.88)	6.6 (3.11)

There were no significant difference in erroneous enumeration rates between males and females (see Table 3) across all of the Dress Rehearsal areas. This differs from 1990 results for the nation as a whole in which males had a higher erroneous enumeration rate than females (Moriarity, 1993).

Table 3: Percentage of Erroneous Enumerations (Standard Error) by Sex

Site	Male	Female
Sacramento	10.8 (0.71)	10.3 (0.74)
Columbia	7.7 (1.07)	7.8 (1.03)
Rural SC	10.7 (1.76)	9.9 (1.98)
Other SC	16.0 (1.61)	17.0 (2.06)
Menominee	9.8 (1.78)	9.7 (1.75)

Age, on the other hand, was related to the erroneous enumeration rate. In the Dress Rehearsal as in the 1990 Census, there were four age poststratification categories: 0-17, 18-29, 30-49, and 50 and over (see Table 4). The Bonferroni adjustment for the four groups (six comparisons) produced an adjusted alpha of 0.017 and the criterion t-value of 2.378.

Across the areas, the 18-29 age group tended to have higher rates of erroneous enumeration than other age groups. In Sacramento, those 18-29 had a higher erroneous enumeration rate than the 30-49 year olds (t=2.895) and those 50 and over (t=5.466). People, in Sacramento, 50 and over had a lower erroneous enumeration rate than 0-17 year olds (t=3.899) and 30-49 year olds (t=4.895). In Columbia, 18-29 year olds also had a higher erroneous enumeration rate than both the 30-49 year olds (2.671) and those 50 and over (t=2.627). In Rural SC, 18-29 year olds had a higher rate of erroneous enumeration than the 0-17 year olds (t=2.590) and the 30-49 year olds (t=2.474). In Other SC, the 18-29 year olds had a higher erroneous enumeration rate than the 39-49 age group (t=2.867). In Menominee we found a different pattern. Those 50 and

older had a higher erroneous enumeration rate than 18-29 year olds (t=2.438). Across the Dress Rehearsal areas, we found no other age categories to be significantly different.

Table 4: Percentage of Erroneous Enumerations (Standard Error) by Age

Site	0-17	18-29	30-49	50+
Sacramento	11.3 (1.06)	12.8 (1.00)	10.7 (0.64)	8.0 (0.63)
Columbia	7.0 (2.31)	9.6 (0.90)	7.6 (1.03)	7.0 (0.58)
Rural SC	9.6 (2.08)	13.8 (2.55)	9.2 (1.64)	10.4 (2.24)
Other SC	18.2 (3.08)	19.7 (2.46)	14.7 (1.44)	15.1 (1.54)
Menominee	8.9 (2.43)	7.4 (3.16)	8.0 (1.84)	13.4 (1.75)

We used different race categories in the various sites reflecting their differing racial makeups. People who marked more than one racial category were assigned to the largest nonwhite category that they marked based on 1990 Census numbers. People who did not mark any racial category were assigned to the non-Hispanic white category. Race/origin groups with less than one percent of the site's 1990 population total were collapsed into the largest nonwhite race according to 1990 data. See Schindler (1999) for a more complete explanation of the racial categories.

Table 5: Percentage of Erroneous Enumerations (Standard Error) in Sacramento by Race

Non-Hispanic White	Non-Hispanic Black	Non-Hispanic Asian	Hispanic
9.0 (0.55)	15.0 (1.32)	10.4 (1.60)	10.6 (0.91)

In Sacramento there were four race categories: non-Hispanic white, non-Hispanic black, non-Hispanic Asian, and Hispanic. All other races were collapsed with Hispanics for estimation purposes. The Bonferroni adjustment for the four groups (six comparisons) produced an adjusted alpha of 0.017 and the criterion t-value of 2.378. The erroneous enumeration rates for the various racial categories in Sacramento ranged from 9.0 percent for non-Hispanic whites to 15.0 percent for non-Hispanic blacks as shown in Table 5. Non-Hispanic blacks had significantly higher erroneous enumeration rate than non-Hispanic whites (t=4.868), non-Hispanic Asians (t=3.450), and Hispanics (4.052). The other racial categories did not differ from each other in their erroneous enumeration rates.

In the South Carolina site there were two racial categories: non-Hispanic white and black. All other race groups were collapsed with blacks for estimation purposes. In Columbia (t=1.333), Rural SC (t=1.010) and Other SC (t=0.708) there were no significant differences between the erroneous enumeration rates of non-Hispanic whites and blacks. See Table 6 for the percentage of erroneous enumerations in each of the three South Carolina areas.

Table 6: Percentage of Erroneous Enumerations (Standard Error) in South Carolina by Race

Site	Non-Hispanic White	Black
Columbia	6.7 (0.62)	8.8 (1.80)
Rural SC	8.9 (1.42)	13.0 (4.12)
Other SC	15.4 (1.53)	17.9 (3.41)

In Menominee there were two racial categories: non-Hispanic white and American Indian. All other race groups were collapsed with American Indians for estimation purposes. As seen in Table 7, American Indians had a 7.4 percent erroneous enumeration rate, while nearly 20 percent of the non-Hispanic white people were erroneously enumerated. The American Indians had a significantly lower erroneous enumeration rate than non-Hispanic whites (t=6.346).

Table 7: Percentage of Erroneous Enumerations (Standard Error) in Menominee by Race

Non-Hispanic White	American Indian
19.8 (1.33)	7.4 (1.47)

When analyzing the erroneous enumeration rates, we considered factors other than the poststrata variables used in the dual system estimator. We also looked at variables related to form such as form length and return type. There were two different form lengths in the Dress Rehearsal: short and long. Approximately 17% of the housing units received a long form which asked for more detailed information about the housing unit and the people living there. The percentage of erroneous enumerations by form type are shown below in Table 8. There was no significant difference in the erroneous enumerations rate of those people who filled out short forms and those who filled out long forms. This is consistent with the 1990 Census results (Griffin and Moriarity, 1992).

Table 8: Percentage of Erroneous Enumerations (Standard Error) by Form Length

Site	Short	Long
Sacramento	10.4 (0.71)	11.0 (1.05)
Columbia	7.6 (1.05)	8.7 (1.19)
Rural SC	9.8 (1.43)	12.5 (4.43)
Other SC	16.4 (1.89)	17.3 (1.92)
Menominee	9.4 (1.64)	12.9 (4.92)

During the Dress Rehearsal, households received a questionnaire which they were supposed to complete and mail back. Those households who did not return their questionnaire were visited by an enumerator who collected the information. For those returns filled out by an enumerator, the information about the household could have come from a household member or it could have been obtained through a proxy interview. For some interviews, the enumerator failed to indicate whether or not the respondent was a proxy. There were four groups, so the adjusted alpha is 0.017 and the criterion t is 2.370. We analyzed Menominee separately due to the small sample size.

We found that mail returns had lower erroneous enumeration rates than both proxy and non-proxy enumerator filled returns. We found that non-proxy enumerator filled returns had lower erroneous enumeration rates than proxy enumerator filled returns. We also found that enumerator filled returns where the proxy information was missing had higher erroneous enumeration rates than mail returns and non-proxy enumerator returns. See Table 9 for the percentage of erroneous enumerations for these variables.

In Sacramento (t=10.411), Columbia (t=3.622), Rural SC (t=3.353) and Other SC (t=2.406) the mail returns had significantly lower erroneous enumeration rates than non-proxy enumerator filled returns.

In Sacramento (t=11.841), Columbia (t=6.281), Rural SC (t=3.315) and Other SC (t=4.657) the mail returns had significantly lower erroneous enumeration rates than proxy enumerator filled returns.

In Sacramento (t=8.979), Columbia (t=5.760), Rural SC (t=2.678) and Other SC (t=3.408) non-proxy enumerator filled returns had significantly lower erroneous enumeration rates than proxy enumerator filled returns.

In Sacramento (t=4.801), Columbia (t=3.303) and Other SC (t=2.882) mail returns had significantly lower erroneous enumeration rates than those enumerator returns where the proxy information was missing. In Rural SC (t=1.570) there was no difference.

In Sacramento (t=2.470), Columbia (t=2.760) and Other SC (t=2.419) non-proxy enumerator filled returns had lower erroneous enumeration rates than those enumerator filled returns where the proxy information was missing. In Rural SC (t=0.977) there was no difference.

In Sacramento (t=3.700) those enumerator filled returns where the proxy information was missing had a lower erroneous enumeration rate than proxy enumerator filled returns.

For Menominee, we only considered mail returns verses enumerator filled returns. Mail returns had 8.8 (1.74) percent erroneous enumerations while enumerator filled returns had 10.3 (3.53) percent erroneous enumerations. There was no difference (t=0.231) between the erroneous enumeration rates of mail returns and those of enumerator filled returns.

Table 9: Percentage of Erroneous Enumerations (Standard Error) by Return Type

Site	Mail	Enumerator		
		No Proxy	Proxy	Missing
Sacramento	6.5 (0.65)	14.1 (0.96)	38.5 (2.77)	22.3 (3.32)
Columbia	6.0 (0.85)	9.0 (1.38)	28.4 (3.68)	22.1 (4.92)
Rural SC	7.0 (1.53)	13.1 (2.56)	33.4 (8.10)	23.5 (10.58)
Other SC	14.1 (1.51)	19.8 (3.14)	34.2 (4.49)	42.4 (10.11)

Next, we looked at the number of people in a household to see if that was related to the erroneous enumeration rate of a person in the household. We compared three groups of households: 1 person, 2-5 people and 6 or more people (see Table 11).

Table 11: Percentage of Erroneous Enumerations (Standard Error) by Number of People in the Household

Site	1 person	2-5 people	6+ people
Sacramento	12.4 (0.88)	10.1 (0.73)	10.9 (1.43)
Columbia	9.2 (0.67)	7.5 (1.21)	7.0 (1.61)
Rural SC	13.9 (2.48)	10.0 (1.76)	8.4 (6.68)
Other SC	18.9 (2.24)	16.3 (1.80)	14.4 (4.74)
Menominee	11.5 (3.89)	10.0 (1.60)	8.2 (4.16)

The Bonferroni adjustment for three comparisons produced an adjusted alpha of 0.035 and a criterion t of

2.114. In Sacramento (t=2.952) we found that people in single person households had a significantly higher erroneous enumeration rate than those people in two to five people households. There were no other differences. This is different from 1990 where it was found that erroneous enumeration rates increased as household size increased (Griffin and Moriarity, 1992).

For those people who did not answer all of the questions on their Census form, the missing values were imputed (see Table 12). It appears that complete data had less error. In Sacramento (t=13.960), Columbia (5.093), Rural SC (t=3.072), and Other SC (3.800) those people with at least one item imputed had significantly higher erroneous enumeration rates that those people with no imputed items. In Menominee (t=0.208) there was no difference.

Table 12: Percentage of Erroneous Enumerations (Standard Error) by

Site	No Imputations	Some Imputations
Sacramento	6.0 (0.57)	16.1 (0.98)
Columbia	5.5 (0.85)	11.1 (1.62)
Rural SC	7.3 (1.55)	13.6 (2.56)
Other SC	13.3 (1.45)	21.7 (2.87)
Menominee	10.1 (1.67)	9.4 (2.89)

IV. Types of Erroneous Enumeration

There are many reasons why a person could be an erroneous enumeration: geocoding errors, fictitious people, duplicate records, other counting errors, insufficient information for matching and unresolved cases.

Geocoding errors occurred when the census placed a housing unit in the wrong block. All of the people in that housing unit were then considered geocoding errors. A fictitious person was another type of erroneous enumeration. A census person could be found to be fictitious if follow-up (after the ICM) determined that the person did not refer to a real person in that block. Duplicates occurred when a person had two or more census records. These additional records were duplicates. Other counting errors occurred when a person was counted in error in a block cluster in the census. The ICM then determined that according to census residency rules the person should have been counted elsewhere. For example, a college student counted in the wrong place or a family with two homes is an other counting error.

People with insufficient information for matching were treated as erroneous enumerations. To have sufficient information for matching, a person had to have had a name and at least one other characteristic provided. People without these two pieces of information were considered to have insufficient information for matching.

An unresolved case occurred when there was not enough information to determine if the person was correctly or erroneously enumerated. A case could have been unresolved because not enough information was collected during the ICM to determine whether or not the person was correctly enumerated during the census. Another example of an unresolved case is a person who was a match, but had an unresolved residency status. Finally, a person who was a possible match, but did not have enough information to positively determine their match status was unresolved. The unresolved cases had their erroneous enumeration probability imputed using a proportion of erroneous enumerations from those people resolved during follow-up. For more information on these categories see Childers (1998).

In Sacramento, geocoding errors and insufficient information for matching were the major causes of erroneous enumeration accounting for about 64 percent of the erroneous enumerations as seen in Table 13.

Table 13: Percentage of Different Types of Erroneous Enumerations (Standard Error) in Sacramento

Type of Error	Percentage
Geocoding error	28.2 (5.08)
Fictitious	10.3 (1.50)
Duplicates	10.3 (1.20)
Other Counting Error	9.5 (1.03)
Insufficient Information	35.4 (2.73)
Unresolved	6.3 (0.66)

In the various South Carolina areas the make-up of the erroneous enumerations varied. In Columbia (see Table 14), insufficient information for matching accounted for approximately 32 percent of the erroneous enumerations. Geocoding errors accounted for about 26 percent of the erroneous enumerations.

Table 14: Percentage of Different Types of Erroneous Enumerations (Standard Error) in Columbia

Type of Error	Percentage
Geocoding Error	26.2 (7.03)
Fictitious	8.1 (1.52)
Duplicates	14.0 (3.36)
Other Counting Error	13.4 (2.01)
Insufficient Information	31.6 (3.61)
Unresolved	6.7 (1.32)

In Rural SC, geocoding errors accounted for about 40 percent of the erroneous enumerations. Duplicates accounted for about 23 percent and other counting errors accounted for about 16 percent of erroneous enumerations. See Table 15 for the different types of erroneous enumerations for Rural SC.

Table 15: Percentage of Different Types of Erroneous Enumerations (Standard Error) in Rural SC

Type of Error	Percentage
Geocoding Error	40.1 (11.97)
Fictitious	9.3 (4.08)
Duplicates	22.9 (6.04)
Other Counting Error	15.7 (4.25)
Insufficient Information	10.2 (2.71)
Unresolved	1.9 (0.61)

In Other SC, geocoding errors accounted for about 64% of the erroneous enumerations (see Table 16). Other SC was update/leave. The different method of enumeration may be the reason for the relatively high percentage of geocoding errors. People with insufficient information for matching accounted for about 12 percent of the erroneous enumerations in Other SC.

Table 16: Percentage of Different Types of Erroneous Enumerations (Standard Error) in Other SC

Type of Error	Percentage
Geocoding Error	63.5 (6.24)
Fictitious	5.7 (2.42)
Duplicates	7.9 (1.67)
Other Counting Error	8.0 (1.46)
Insufficient Information	11.6 (1.75)
Unresolved	3.4 (1.45)

In Menominee, over half (50.8 percent) the erroneous enumerations were due to duplicate records for the same person (see Table 17). Other counting errors contributed to about 32 percent of the erroneous enumerations.

Table 17: Percentage of Different Types of Erroneous Enumerations (Standard Error) in Menominee

Type of Error	Percentage
Geocoding Error	1.3 (1.38)
Fictitious	1.3 (1.00)
Duplicates	50.8 (8.82)
Other Counting Error	31.5 (10.40)
Insufficient Information	13.5 (6.32)
Unresolved	1.6 (0.85)

V. Conclusions

The percentage of erroneous enumerations varied across the sites ranging from 7.7 percent in Columbia to 16.5 percent in Other SC. We determined erroneous enumeration rates for various subgroups of the population. We looked at the poststrata categories, form characteristics and other variables.

In general, we found that renters had higher erroneous enumeration rates than owner and 18-29 year olds had higher erroneous enumeration rates than other age groups. We found no differences in the erroneous enumeration rates between the sexes.

Form length was not related to erroneous enumeration rates. However, return type was related to erroneous enumeration rates. Mail returns had lower erroneous enumeration rates than enumerator filled returns. For enumerator filled returns, non-proxy responses had lower erroneous enumeration rates than proxy responses. We also found that those people with some variables imputed had higher erroneous enumeration rates than those people with no variables imputed. Proxy responses and imputed values are an indicator of poor data quality. Although these results are only representative of the Dress Rehearsal sites, they do show that the quality of the data is related to the erroneous enumeration rate. It is difficult to draw conclusions about the 2000 Census from these data because the census methods are not the same as in the Dress Rehearsal.

VI. References

Childers, Danny (1998), "The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement," *DSSD Census 2000 Dress Rehearsal Memorandum Series, Chapter F-DT-2.*

Games, Paul (1971) "Multiple Comparisons of Means," *American Educational Research Journal*, 8, 3, 531-565.

Griffin, Deborah and Moriarity, Christopher (1992) "Characteristics of Census Errors," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 512-517.

Moriarity, Christopher (1993), "Characteristics of Census Error – Additional Results," *1990 Decennial Census Preliminary Research and Evaluation Memorandum Number 240.*

Schindler, Eric (1999) "Iterative Proportional Fitting in the Census 2000 Dress Rehearsal," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, to appear.