Accuracy and Coverage Evaluation Persons Not Matched in Census 2000

Glenn Wolfgang, Peter P. Davis, Phawn Stallone, Bureau of the Census
Glenn Wolfgang, Bureau of the Census, Washington, DC 20233

Key Words: Census Coverage; Dual System Estimate

1. Introduction

The Accuracy and Coverage Evaluation (A.C.E.) involved two samples. Both were selected in sample areas, consisting of approximately 300,000 housing units in 11,303 block clusters in the fifty states and the District of Columbia. People enumerated by the census in A.C.E. sample areas made up the E sample, which was used to measure errors among census enumerations. Census Day residents in sample areas listed by the A.C.E. survey made up the P sample, which was used to determine who was missed in the census. Names and characteristics of P-sample people were compared to those of census enumerations in the sample cluster or designated surrounding blocks. Matches were persons found in both, that is, a record existed for them among both A.C.E. and Census 2000 enumerations.

The focus of this analysis was on P-sample nonmatches. It addressed only the undercount aspects of census coverage evaluation. The aim was to identify characteristics that were related to being missed in census enumeration. The statistic used in this study was the percent not matched, the percent of nonmatches among P-sample persons, computed within age, race, or other descriptive variables. The proportion not matched (NM/P), like the number of data-defined census persons excluding whole-person imputations or additions to the census too late to be included in matching (DD) and the proportion of correct enumerations as determined by the E sample (CE/E), has an important role in the dual system estimate (DSE) formula:

$$DSE = (DD) * (CE/E) / (1 - (NM/P)).$$

There were many ways to investigate the nonmatches. The major approach in this report was to divide the P sample into groups on the basis of levels of important variables, compute a percent not matched for each level and test for differences. Percentages not matched have been studied independently of the effects of erroneous enumerations. Prior P-sample nonmatch analyses by Hogan (1993), Moriarity and Childers (1993), and Wolfgang and Childers (1999) provided comparable results cited in this report. Erroneous enumerations in Census 2000 were investigated by Feldpausch (2001). Beaghen, Feldpausch, and Byrne (2001) modeled both E-sample and P-sample data to gain insight into missed enumerations, but are beyond the scope of this writing.

Other prior publications provided general background to this research. Hogan (1993) reported on both analyses and procedures for the 1990 census. Hogan (2000) described application of theory in A.C.E. Childers (2001) described the A.C.E. design. Adams, Barrett, and Byrne (2001) summarize procedures for A.C.E. operations.

2. Methods

This study used the person-level records of Census 2000 and of the independent A.C.E. enumeration. P-sample person records and census person records were computer matched within cluster. The computer matching involved first standardizing the name formats. Next, names and person characteristics of the P-sample people were compared to those of census people with sufficient information for matching and follow-up. A ranking score was assigned to each pair of person records and the optimal pairings were identified. Those pairs were reviewed and the scores used to separate matches from possible matches and from nonmatches. Score cutoffs identifying matches were assigned conservatively to minimize the number of false matches.

The possible matches and nonmatches in the P-sample were clerically reviewed using an automated match and review system. The names, age, race, Hispanic origin, sex, relationship, household composition, and address were displayed for review by the matching clerks, who matched some people the computer could not. After the matching, field follow-up was conducted to confirm or resolve who matched and who should have been counted in the cluster on Census Day.

Final dual system estimates were weighted or adjusted for the results of various A.C.E. operations. In addition to initial sampling, large clusters with eighty or more housing units in a block were subsampled within the

The authors are mathematical statisticians in the Decennial Statistical Studies Division of the U.S. Census Bureau. 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.

block to reduce the intra-cluster correlation and to reduce the interviewing workloads and given an additional subsample weight. A Targeted Extended Search operation identified potential incorrect assignment of the block cluster identity code and extended search for relevant person data. It improved the precision of estimates and improved the robustness of dual system estimates. TES involved sampling -- and corresponding weights. See Wolfgang, Stallone, and Adams (2001) for more information and analyses of the TES. In addition, if match status remained unresolved, match probability was imputed. If residence status remained unresolved, residence probability was imputed. Missing values for post-stratification variables, namely tenure, age, sex, race, and Hispanic origin, were imputed. For households not successfully interviewed, a non-interview adjustment was applied. Match probabilities, residence probabilities, and final sampling weights (incorporating all these operations including TES selection and non-interview adjustments) were applied in all analyses.

The percent not matched is the statistic analyzed in this work; it is a percentage form of the nonmatch rate. It is the weighted number of P-sample nonmatches divided by the weighted number of P-sample persons expressed as a percent. It was computed for various groups within the P sample. Identifying groups with unusually high rates of nonmatches provided insights into conditions associated with missed census enumerations. For this purpose, P-sample persons were grouped into meaningful levels of a variety of variables, especially variables used for post-stratification (Haines, 2001) and others that were expected to be related to the percent of nonmatches.

The P sample analyzed in this report included nonmover and outmover data. See the discussion in the Limitations section regarding how official estimates use inmover data as well. Most of the analyses were done using the whole P sample. Those analyses were conducted using variables from A.C.E. data collection or processing.

A percent not matched was computed for different P-sample subgroups defined by levels of a variable's values. Stratified Jackknife methods were used to compute variance estimates for the percentages not matched. Then, t values were produced for paired comparisons of the rates. Statistical significance for each t value was determined using the Bonferroni multiple comparison of means technique, which controls the probability of Type I error for a family of tests. In the context of this analysis, a family of tests was defined as all tests conducted among sample subgroups formed from the variable under analysis. For example, when comparing four subgroups, six pairs of statistics were tested. To control the chance of Type I error at $\alpha = 0.10$ for all six tests combined, we used an adjusted criterion t-value associated with the probability of one of six two-tailed tests that had a joint error probability equal to 0.10. In addition, tests with levels based on less than 100 person records were avoided, either through collapsing with other levels or simply by dropping the level from that family of tests.

3. Limits

This analysis of A.C.E. data had certain research limits. It had a specific focus on P-sample nonmatches. It did not address the impact of other total error components (Mulry and Spencer, 1991), even errors in collected census data measured by the E sample.

Nonmatch statistics in this analysis were different from the official statistics computed during production. Nonmatch statistics in this study were computed solely using nonmovers and outmovers; inmovers were not used. Official nonmatch statistics were computed using a combination of nonmover, outmover, and inmover information. For official dual system estimation, statistics were computed and defined for levels of post-stratum variables. In these analyses, we were interested in some non-post-stratum variables and used the simpler methodology. Haines (2001) and Davis (2001) elaborated on the different methodologies for handling movers. This analysis procedure yielded percentages not matched that were a little lower than official percentages not matched, typically about 0.3 percent within major population subgroups, as seen in Davis (2001). These small, fairly consistent differences were not expected to affect any of the significance tests reported here.

Variance computations in these analyses were simplified and did not take all levels of the sampling into account. We expected only trivial impacts on variances due to variance computations; we expected no impact on test results and conclusions.

4. Results

The overall percent not matched, 8.2 percent, was computed from a weighted total of about 21,150,000 nonmatches and a weighted total of about 258,550,000 P-sample persons. The percentages reported below ranged from 5.3 percent (among spouses of the first person listed in the A.C.E. interview) to around 22.6 percent (among outmovers).

Results from comparing statistics are presented below in tables displaying variable level names with level numbers assigned for reference in another table column, values for the statistic of interest (in columns headed "percent"), a list of the level numbers with which a significant difference was found (in columns headed "differs from"), the stratified jackknife standard error (in columns headed "s.e."), and the weighted percent of persons contributing data to the analysis (in columns headed "n"). The criterion t value that applies in that table is noted below each table. The levels are arranged in ascending order of percent not matched to help display data patterns.

The major analyses' results are presented below. Of primary interest were variables used to form post-strata in estimating dual system estimates Haines (2000). They were analyzed using the levels as defined for post-stratification.

Table 1 shows that percentages not matched differed by age and sex post-strata levels except in one comparison involving children. Generally, from younger to older adults, the percent not matched decreased, with males generally having higher rates at each age. Children's rates were close to the median of groups aged 18 to 49, commonly child-raising ages. We might speculate that the child's nonmatch rates relate to their parents' ages.

Table 1: Percent Not Matched by Age and Sex

Age, Sex	Per-cent	Differs from	s.e.	n
1: 50+ Female	5.6	all	0.1	14.9
2: 50+ Male	6.2	all	0.2	12.3
3: 30-49 Female	6.9	all	0.1	16.2
4: 30-49 Male	8.5	1,2,3,6,7	0.2	15.2
5: 0-17	8.8	1,2,3,6,7	0.2	26.2
6: 18-29 Female	11.1	all	0.2	7.7
7: 18-29 Male	13.2	all	0.3	7.5

Note: Criterion for levels to differ was | t | > 2.815

Typically, home owners had much lower nonmatch rates than non-owners, as Table 2 results show.

Table 2: Percent Not Matched by Home Ownership

Tenure	Per-cent	Differs from	s.e.	n
1: Owner	6.1	2	0.1	69.8
2: Non-owner	13.1	1	0.2	30.2

Note: Criterion for levels to differ was | t | > 1.645

Respondents to Census 2000 were able to self-identify with more than one race group. Combining 63 levels of Race with two levels of Hispanic Origin yielded 126 possible Race/Hispanic Origin groups. Rules were adopted to assign persons in those 126 groups to one of seven Race/Hispanic Origin Domains (See Haines, 2001).

Table 3 shows that Hispanic, Non-Hispanic Black, American Indian on Reservation, American Indian off Reservation, and Native Hawaiians or Pacific Islander Domains had higher percentages not matched than the Non-Hispanic White or "Some other race" Domain. The Non-Hispanic Asian Domain had a higher percent not matched than the Non-Hispanic White or "Some other race" Domain and had a lower percent than three other levels. Other differences were not significant.

Table 3: Percent Not Matched by Race and Hispanic Origin

Race and Hispanic Origin Domain	Per-cent	Differs from	s.e.	n
1: Non-Hispanic White or "Some other race"	6.7	all	0.1	72.1
2: Non-Hispanic Asian	9.2	1,4,5,6	0.5	3.4
3: American Indian off Reservation	11.7	1	1.1	0.5
4: Hispanic	12.1	1,2	0.3	12.3
5: Non-Hispanic Black	12.8	1,2	0.3	11.4
6: American Indian on Reservation	13.7	1,2	1.1	0.2
7: Native Hawaiian or Pacific Islander	15.0	1	2.5	0.2

Note: Criterion for levels to differ was | t | > 2.815

Metropolitan Statistical Area (MSA) and Type of Enumeration Area (TEA) were combined to form one variable used in post-stratification. MSAs denoted the boundaries of cities or other areas named for statistical purposes. Most of the population was in the Mailout/Mailback TEA, in which people receive and return census forms by mail. Mailout/Mailback areas were divided into three levels based on size of the MSA. A fourth level was comprised of other areas where census workers visited to list or update addresses or conduct enumerations on the spot. Although MSA/TEA was used to post-stratify only Hispanic, Non-Hispanic Black, and Non-Hispanic White or "Some other race" Domains, this analysis included values for all P-sample persons.

Table 4 shows that the extremes, large MSAs and areas where enumeration was not conducted by mail, had higher percentages not matched. Perhaps the most urban and the most rural areas have different causes (possibly mobility for large MSAs and inaccessibility for very rural areas) for being harder to enumerate than the more developed rural and suburban areas that had unique postal addresses.

Table 4: Percent Not Matched by Size of Metropolitan Statistical Area and Type of Enumeration Area

MSA/TEA	Per-cent	Differs from	s.e.	n
1: Small MSA & Non-MSA Mailout/Mailback	7.3	3,4	0.2	20.2
2: Medium MSA Mailout/Mailback	7.4	3,4	0.2	31.3
3: Large MSA, Mailout/Mailback	9.0	1,2	0.2	30.4
4: All Other TEAs	9.2	1,2	0.3	18.1

Note: Criterion for levels to differ was | t | > 2.386

The tract-level return rate, a sign of public cooperation, was the proportion of occupied housing units in a census tract that returned a 2000 census questionnaire. High and low return rate indicator values were assigned for the Non-Hispanic White or "Some other race," Non-Hispanic Black, and Hispanic domains. 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 (Haines, 2001).

Table 5 shows that persons in high return rate post-strata had a lower percent not matched than other P-sample persons.

Region of the U.S. was also used to post-stratify homeowners who were Non-Hispanic White or "Some other race". In an analysis of the whole P sample, the Midwest region stood out with a lower nonmatch rate than other areas, as shown in Table 6.

Age, sex, race, Hispanic origin, and tenure were sometimes imputed for the A.C.E. For persons without imputation of post-stratum characteristics, nonmatch rates were lower, as shown in Table 7.

Table 5: Percent Not Matched by Census Return Rate Indicator

Return Rate Indicator	Per-cent	Differs from	s.e.	n
1: High	7.1	all	0.1	72.3
2: Not Applicable	10.0	all	0.5	4.3
3: Low	11.1	all	0.3	23.4

Note: Criterion for levels to differ was | t | > 2.121

Table 6: Percent Not Matched by Region of the United States

Region	Per-cent	Differs from	s.e.	n
1: Midwest	6.1	all	0.2	22.9
2: Northeast	8.3	1	0.3	19.0
3: West	8.7	1	0.3	22.8
4: South	9.1	1	0.2	35.3

Note: Criterion for levels to differ was | t | > 2.386

Table 7: Percent Not Matched by Imputation of Characteristics

Imputed or Not	Per-cent	Differs from	s.e.	n
1: Not Imputed	7.9	2	0.1	94.7
2: Imputed	13.8	1	0.4	5.3

Note: Criterion for levels to differ was | t | > 1.645

Table 8 shows that subsampled clusters had a higher percent not matched.

Table 8: Percent Not Matched by Involvement in Subsampling

Subsampled or Not	Per-cent	Differs from	s.e.	n
1: Not Subsampled	7.8	2	0.1	63.5
2: Subsampled	8.8	1	0.2	36.5

Note: Criterion for levels to differ was | t | > 1.645

Table 9 shows that movers in 2000 had a higher percent not matched than nonmovers. Outmovers, Census Day residents who moved from the sample address before the survey interview, represented movers in 2000 data. In 1990 data analyses conducted by Moriarity and Childers (1993), movers, estimated from inmovers, also had a higher rate (24.8 percent). Wolfgang and Childers (1999) reported a higher rate for movers in all four census dress rehearsal sites.

Table 9: Percent Not Matched by Mover Status

Person Mover Status	Per-cent	Differs from	s.e.	n
1: Nonmover	7.7	2	0.1	96.6
2: Outmover	22.6	1	0.5	3.4

Note: Criterion for levels to differ was | t | > 1.645

When the respondent was a member of the household and not a proxy, the percent not matched was lower, as shown in Table 10.

Table 10: Percent Not Matched by Type of Respondent

Person Mover Status	Per-cent	Differs from	s.e.	n
1: Household Member	7.5	2	0.1	94.5
2: Proxy	20.3	1	0.4	5.5

Note: Criterion for levels to differ was | t | > 1.645

Other analyses of this type were done but their tables are not included here. Some characteristics of the household or of persons appeared likely to relate to nonmatch rates, including the type of structure at the basic address, household size, and how closely related the person is to a central person in the household.

A significantly lower percent not matched was found among residents of single family permanent structures (6.5 percent), relative to mobile homes (13.3 percent), multi-unit structures like apartment buildings (13.4 percent), or miscellaneous other structures like living quarters within a special place (22.1 percent).

Household size was based on the number of P-sample persons listed at the address. Up to six persons could be enumerated on most census forms; additional forms or procedures were usually needed to enumerate seven or more. Large households (in this case, those with seven or more residents – generally requiring additional census questionnaires or forms) had a higher percent not matched (17.2 percent). Households with only one reported resident had a higher percent not matched than those with a few other residents. If solitary residents were more likely to be mobile, mobility could be related to percent not matched.

Kinship was a measure of how closely related the person was to the central person in the household. The respondent for A.C.E. 2000 designated someone in the household, usually the person in whose name the residence was owned or rented, to be listed first in the data collection. The question, "What is . . . 's relationship to . . . ?" was asked about subsequent persons listed and referred to the first person listed. Kinship for other persons in the household was in relationship to this reference person. Reference persons either lived alone or shared the housing unit with other relatives or nonrelatives.

The less closely related a person was to the reference person, the higher the percent not matched. Reference persons living alone had a higher percent not matched (10.2 percent) than those living with others (6.2 percent). A spouse's percent not matched was the lowest (5.3 percent). The parent/child group had a moderate rate (8.0 percent). Other relatives and nonrelatives had the highest (17.0 percent). Perhaps kinship categories reflected mobility or how likely household members were to move.

Table 11 shows percentages not matched for different types of nonmatches. Nonmatch type was defined by whether others in the household were matched to census persons and by whether the housing unit address was matched to a census address. Specifically, nonmatch type was based on three categories of whole-household match status: partial-household nonmatches, whole-household nonmatches in matched housing units, and whole-household nonmatches in nonmatched housing units.

The nonmatch type analysis differed in two ways from analyses reported above. First, it included only P-sample nonmovers that were resolved (that is, without imputation of residence or match status). Second, data from 2000 and from 1990 (Hogan, 1993) are presented as a percent not matched out of the total of resolved P-sample nonmovers, rather than the number of P-sample cases within the subgroup. As a consequence, the percentages add up to overall resolved nonmover percentages not matched (7.2 percent for 2000 and 5.9 percent for 1990).

The percent of nonmatches among nonmovers increased from 1990 to 2000. Increased housing unit matches may have been related to that increase. The increase could also be related to an increase of data that were not available for matching: census cases held back in processing while they were confirmed to be not duplicated and later reinstated, whole-person imputations, and person records with insufficient information for matching and follow-up.

Table 11: Nonmatch Household Types: Percent of Resolved Nonmovers Not Matched for 2000 and 1990

Nonmatched Person by Household Type	2000		1990	
	Percent of Nonmovers	Percent of Nonmatches	Percent of Nonmovers	Percent of Nonmatches
Partial-Household Nonmatch	2.2	30.0	1.8	30.4
Whole-Household Nonmatch in Matched Housing Unit	3.3	45.9	2.3	38.5
Whole-Household Nonmatch in Nonmatched Housing Unit	1.7	24.1	1.8	31.1
TOTAL	**7.2**	**100**	**5.9**	**100**

Note: Movers and unresolved cases were removed from both 1990 and 2000 analyses.

5. Conclusions

The persistence of differences between post-strata levels confirmed the importance and validity of those levels.

Other variables were found to be related to percentages not matched. High percentages not matched (higher than 10 percent) were associated with: outmovers (22.6 percent), proxy respondents (20.3 percent), residents in structures other than a single-family dwelling (13.3 percent to 22.1 percent), seven or more residents at the address (17.2 percent), distant or no kinship (i.e., not parent, child, or spouse) to person listed first on the questionnaire (17 percent), imputed post-stratification variables (13.8 percent), non-owners (13.1 percent), young adults, age 18 to 29 (11.1 percent to 13.2 percent), native Hawaiian or Pacific Islander, American Indian, Blacks, Hispanics (11.7 percent to 15 percent), and low census return rate indicator (11.1 percent).

6. References

Adams, T., Barrett, D., and Byrne, R. (2001). "Operational Plan for Accuracy and Coverage Evaluation (A.C.E.) for Census 2000," DSSD Census 2000 Procedures and Operations Memorandum Series S-TL-06, U.S. Census Bureau, Washington, D.C.

Beaghen, M., Feldpausch, R., and Byrne, R. (2001). "Modeling Accuracy and Coverage Evaluation Nonmatches in the Census 2000," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.

Childers, D. (2001). "The Design of the Census 2000 Accuracy and Coverage Evaluation (A.C.E.)" DSSD Census 2000 Procedures and Operations Memorandum Series S-DT-01, U.S. Census Bureau, Washington, D.C.

Davis, P. (2001). "Accuracy and Coverage Evaluation: Dual System Estimation Results," DSSD Census 2000 Procedures and Operations Memorandum Series B-9*, U.S. Census Bureau, Washington, D.C.

Feldpausch, R. (2001). "Census 2000 E-Sample Erroneous Enumerations," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.

Hogan, H. (1993). "The 1990 Post-Enumeration Survey: Operations and Results," Journal of the American Statistical Association, 88, 1047-1060.

Hogan, H. (2000). "Accuracy and Coverage Evaluation: Theory and Application," Internal document, U.S. Census Bureau, Washington, D.C.

Haines, D. (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 Q-48, U.S. Census Bureau, Washington, D.C.

Moriarity, C. and Childers, D. (1993). "Analysis of Census Omissions: Preliminary Results," Proceedings of the Section on Survey Research Methods, American Statistical Association, 629-634.

Mulry, M. and Spencer, B. (1991). "Total Error in PES Estimates of Population," Journal of the American Statistical Association, 86, 839-854.

Wolfgang, G. and Childers, D. (1999). "Integrated Coverage Measurement Persons Not Matched in the Census 2000 Dress Rehearsal," Proceedings of the Section on Survey Research Methods, American Statistical Association, 725-730.

Wolfgang, G. and Stallone, P. (2001). "Targeted Extended Search in the Accuracy and Coverage Evaluation of the Census 2000," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.