# COVERAGE MEASUREMENT RESULTS FROM THE CENSUS 2000 ACCURACY AND COVERAGE EVALUATION SURVEY

**Dawn E. Haines and Peter P. Davis**
**Dawn E. Haines, Bureau of the Census, Washington, DC 20233**

KEY WORDS: Post-stratification, Dual System Estimation, Coverage Correction Factor

## I. Introduction

The Census 2000 Accuracy and Coverage Evaluation (A.C.E.) Survey was designed to measure the coverage properties of Census 2000. Post-strata were defined to reduce the heterogeneity in the population as much as possible without substantially increasing the variance of individual post-strata. The post-stratification plan used the variables race/Hispanic origin domain, age/sex, tenure, Metropolitan Statistical Area/Type of Enumeration Area, return rate, and census region to form a maximum number of 448 post-strata. Dual system estimates (DSEs) were computed in order to provide population estimates by post-strata. Coverage correction factors (CCFs) were then computed as the ratio of the DSE to the census count for that post-stratum. This paper compares the coverage patterns of subpopulations defined by the post-stratification variables. In addition, specific DSE components such as mover match rates and correct enumeration rates are presented.

The Accuracy and Coverage Evaluation (A.C.E.) Survey relies on dual system estimation to estimate coverage in Census 2000. The Census Bureau obtains a roster from the A.C.E. block clusters independently of the census. The independent roster (P Sample) and the census roster (E Sample) are matched; the results of the matching and followup interviewing are used to estimate the total number of persons in the census. These estimates reflect the coverage of the census, either a net undercount or a net overcount. Estimates are calculated separately within population subgroups called post-strata. Post-stratum

_____

estimates are then used to determine coverage correction factors which are applied to all people counted in the census, according to their assigned post-stratum.

This paper documents the Census 2000 A.C.E. dual system estimation results for the U.S. The tables highlight the percent net undercount for the major demographic groups and summarize the DSE components.

## II. Methodology

The dual system estimate (DSE) is a population size estimator while the coverage correction factor (CCF) and the percent net undercount (UC) are coverage estimates. For a given post-stratum, the dual system estimate is defined as follows:

$$ DSE = DD \times \frac{CE}{N_e} \times \frac{N_p}{M} $$

where

| | | |
|---|---|---|
| DD | = | the number of census data-defined persons eligible and available for A.C.E. matching; |
| CE | = | the estimated number of correct enumerations from the E Sample; |
| $N_e$ | = | the estimated number of people from the E Sample; |
| $N_p$ | = | the estimated total population from the P Sample; |
| M | = | the estimated number of persons from the P-Sample population who match to the Census. |

The CCF is a measure of correction to assess the degree of net overcount or net undercount of the household population within the Census. The CCF for a post-stratum is the ratio of the DSE to the census count for that post-stratum, written as

$$ CCF = \frac{DSE}{C} $$

where

| C | = | the final census household population count where C=DD+II+LA; |
| --- | --- | --- |
| II | = | the number of census people with insufficient information; |
| LA | = | the number of people added (late) to the census and not available for A.C.E. matching. Late Adds include both data-defined and non-data-defined records. |

Coverage correction factors are primarily used to form synthetic estimates. For example, a CCF of 1.05 implies that for every 100 person records within a given post-stratum, the net undercount is five persons. On the other hand, for every 100 person records within a particular post-stratum, a CCF of 0.95 implies a negative net undercount, or a net overcount, of five persons.

The percent net undercount (UC) is the estimated net undercount (or net overcount) divided by the dual system estimate for a post-stratum, expressed as a percentage. A positive number implies undercoverage while a negative number implies overcoverage. The percent net undercount for Census 2000 in this paper is strictly for the household population and excludes Group Quarters persons. Therefore,

$$UC = \left( \frac{DSE - C}{DSE} \right) \times 100.$$

## III. Post-Stratification

The goal of post-stratification is to group individuals with similar census inclusion probabilities together. Logistic regression modeling was used on the 1990 Post Enumeration Survey (PES) data to determine the best indicators of capture in the census. This work is discussed in Haines and Hill (1998). DSEs were calculated within post-strata to reduce heterogeneity bias while maintaining acceptable post-stratum variances. Haines (2000) documents the Census 2000 A.C.E. post-stratification design while Haines (2001) presents detailed specifications for computing DSEs.

The variables race, Hispanic origin, age, sex, tenure, Metropolitan Statistical Area, type of enumeration area, return rate, and census region define the post-strata. There are 64 post-stratum groups with each containing seven age/sex categories. This results in a maximum number of $64 \times 7 = 448$ post-strata. The 448 post-strata were pre-collapsed based on expected sample sizes. Further collapsing patterns were pre-specified for cells with small P-Sample sizes and outlier coefficients of variation (CVs). Post-collapsing due to small sample sizes or outlier CVs results in fewer than 448 post-strata. A post-stratum is deemed too small if the sum of the nonmover and outmover sample sizes is less than 100.

For the 2000 A.C.E., the minimum sample size requirement was not realized seven times while the outlier CV condition occurred once. For these eight post-stratum groups, the pre-specified collapsing rules require that the seven age/sex groups be collapsed into three categories: 0-17, 18+ Male, and 18+ Female. No further collapsing was required since the remaining 416 post-strata satisfied both constraints. As a result, the final 2000 A.C.E. post-stratum design contains 416 direct dual system estimates.

## IV. Changes Since 1990

This section highlights some of the major differences between the 1990 Post Enumeration Survey and the Census 2000 Accuracy and Coverage Evaluation Survey. It's helpful to recall these points when contrasting the 1990 PES and 2000 A.C.E. results in Table 1.

## A. Multiple Race

Multiple race reporting was allowed for the first time in Census 2000. For post-stratification purposes, the 63 race and two Hispanic origin categories were combined into seven race/Hispanic origin domains. Specific rules for defining these domains, especially for persons with multiple race responses, are found in Haines (2001). For example, a person responding to the census as Black, Asian, and Non-Hispanic would be assigned to the Non-Hispanic Black domain. The seven race/Hispanic origin domains are:

- Non-Hispanic White or "Some other race"
- Non-Hispanic Black
- Hispanic
- Native Hawaiian or Pacific Islander
- Non-Hispanic Asian
- American Indian or Alaska Native on Reservation
- American Indian or Alaska Native off Reservation

In contrast, the 1990 PES had five race/origin domains based on single-race reporting. They are:

- Non-Hispanic White & Other (including American Indian off Reservation)
- Black
- Hispanic White & Other
- Asian & Pacific Islander
- American Indian on Reservation (including Alaska Native)

## B. Universe

The Census 2000 A.C.E. estimates presented in this paper are for the household population excluding persons in the Remote Alaska type of enumeration area. Group Quarters persons are excluded from the Census 2000 A.C.E. universe since this population is mobile and much less likely to be enumerated in the census and the P Sample in the same location. In contrast, the 1990 PES estimates include some non-institutional Groups Quarters such as college dormitories. All other features of the universes are the same.

## C. Treatment of Movers

Some persons will move between Census Day and A.C.E. Interview Day. A mover is a person whose housing unit on A.C.E. Interview Day differs from that on Census Day. The 2000 A.C.E. treats movers by Procedure C (PES-C). This procedure identifies all residents living or staying in the housing unit at the time of the A.C.E. interview (nonmovers and inmovers). In addition, all other persons who lived in the housing unit on Census Day who have since moved (outmovers) are identified. For outmovers, a proxy interview is attempted in order to obtain data such as name, sex, and age which is used for matching. The mover match rate is obtained using outmover match rates. On the other hand, the total number of movers is estimated using inmovers. No matching is conducted for inmovers.

If the outmover sample size in a post-stratum is less than 10, movers are treated using Procedure A (PES-A). This procedure uses outmover information to estimate both the mover match rate and the number of movers. For the 2000 A.C.E., Procedure A was implemented 63 times out of a possible 416 post-strata. Individual DSE components under Procedures A and C are defined in Haines (2001).

The 1990 PES used Procedure B (PES-B). This procedure identifies all residents living or staying in the housing unit at the time of the PES interview. The respondent is asked to provide the address(es) where these persons were living or staying on Census Day. These persons are then matched, based on their Census Day address.

## V. Results

Coverage results are given at the national level and for major demographic groups. All 1990 and 2000 estimates are based on direct DSEs using estimation definitions. Percent net undercount estimates and their standard errors are presented for the 2000 A.C.E. and 1990 PES data. The standard errors are computed using the methodology given in Kim et al. (2000) and Navarro and Sands (2001). Comparisons between the 2000 and 1990 estimates are made when applicable. Summaries of the DSE components are presented for the A.C.E. data. See Davis (2001) for a more comprehensive summary of results, including DSE component estimates, sample sizes, and variances.

## A. Percent Net Undercount

Table 1 presents the percent net undercount and their standard errors for major demographic groups in the 2000 A.C.E. and the 1990 PES. Dual system estimation shows that Census 2000 undercounted the national household population and differentially undercounted population subgroups. Relative to the 1990 census, Census 2000 showed improvement in the overall percent net undercount and the differential undercounts of certain population groups. The national net undercount of the household population for Census 2000 is 1.18 percent. For the 1990 census, the national net undercount was 1.61 percent. (Recall that the 1990 PES universe is defined differently than the 2000 A.C.E. universe.)

Census 2000 coverage patterns show differential undercount rates among the race/Hispanic origin domains, tenure groups, and the age/sex categories. For the race/Hispanic origin domains, the percent net undercount ranges from 0.67 percent for Non-Hispanic White or "Some other race" to 4.74 percent for the American Indian On Reservation domain. For the 1990 census, the net undercount ranged from 0.68 percent for Non-Hispanic White & Other to 12.22 percent for the American Indian on Reservation domain.

The standard errors fell for all directly comparable race/Hispanic origin domains. This reduction is seen most clearly for the American Indian On Reservation domain. The lower standard error for this domain could be due to a change in census methodology for American Indian reservations (List Enumerate in 1990 to Update Leave in 2000) and the fact that this population is oversampled. The net undercount rates for the Non-Hispanic Black and Hispanic domains are 2.17 and 2.85 percent, respectively. In 1990, the corresponding net undercount rates were 4.57 and 4.99 percent, showing an approximate 50 percent reduction in the net undercount rate for these two domains. The 2000 net undercount rates for the Non-Hispanic Black and Hispanic domains are not significantly different at the $\alpha = 0.10$ level.

## Table 1. Percent Net Undercount for Major Groups: 2000 A.C.E. and 1990 PES

| | 2000 A.C.E. | | 1990 PES | | |
| --- | --- | --- | --- | --- | --- |
| Characteristic | Net Undercount (%) | Standard Error (%) | Net Undercount (%) | Standard Error (%) | Characteristic |
| Total | 1.18 | 0.13 | 1.61 | 0.20 | Total |
| Race/Origin Domain | | | | | Race/Origin Domain |
| Non-Hispanic White | 0.67 | 0.14 | 0.68 | 0.22 | Non-Hispanic White & Other |
| AI Off Reservation | 3.28 | 1.33 | | | |
| Non-Hispanic Black | 2.17 | 0.35 | 4.57 | 0.55 | Black |
| Hispanic | 2.85 | 0.38 | 4.99 | 0.82 | Hispanic |
| Non-Hispanic Asian | 0.96 | 0.64 | 2.36 | 1.39 | Asian or Pacific Isl |
| Hawaiian or Pacific Isl | 4.60 | 2.77 | 12.22 | 5.29 | AI On Reservation |
| AI On Reservation | 4.74 | 1.20 | | | |
| Tenure | | | | | Tenure |
| Owner | 0.44 | 0.14 | 0.04 | 0.21 | Owner |
| Non-Owner | 2.75 | 0.26 | 4.51 | 0.43 | Non-Owner |
| Age/Sex | | | | | Age/Sex |
| 0-17 | 1.54 | 0.19 | 3.18 | 0.29 | 0-17 |
| 18-29 Male | 3.77 | 0.32 | 3.30 | 0.54 | 18-29 Male |
| 18-29 Female | 2.23 | 0.29 | 2.83 | 0.47 | 18-29 Female |
| 30-49 Male | 1.86 | 0.19 | 1.89 | 0.32 | 30-49 Male |
| 30-49 Female | 0.96 | 0.17 | 0.88 | 0.25 | 30-49 Female |
| 50+ Male | -0.25 | 0.18 | -0.59 | 0.34 | 50+ Male |
| 50+ Female | -0.79 | 0.17 | -1.24 | 0.29 | 50+ Female |

2000 net undercount is for household population.
1990 net undercount is for the PES universe which included noninstitutional Group Quarters in addition to the household population. As a result, the 1990 estimates may differ from the Committee on Adjustment of Postcensal Estimates (CAPE) results. See Bryant et al. (1992) and Thompson (1992).
The 1990 Hispanic domain excludes Blacks, Asian or Pacific Islanders, and American Indians on Reservation.
A negative net undercount denotes a net overcount.

Tenure is an important indicator of census coverage. Non-owners were counted much better in Census 2000 relative to 1990. This is reflected by a net undercount of 2.75 percent in 2000 as compared to 4.51 percent in 1990.

The coverage of children improved. In 1990, their net undercount was 3.18 percent. This figure dropped to 1.54 percent in 2000. As shown in Table 1, standard errors for all age/sex groups in 2000 were lower than their 1990 levels. Based on a two-sided hypothesis test with $\alpha$ = 0.10, the percent net undercount for Males ages 18 to 29 years is higher than the other six age/sex groups. Also, the percent net undercount for Females ages 50 or older is lower than the other six age/sex groups. Males and females who are 50 years or older have negative net undercount rates, denoting net overcounts.

The sampling variance was expected to be lower in 2000 relative to 1990 for several reasons. First of all, the housing unit sample size for the 2000 A.C.E. was almost double that of the 1990 PES. In 2000, better measures of population size were available during cluster sample selection. Finally, sampling weights were less variable. For large geographic areas, the actual reduction in sampling variance was typically greater than the 25 percent reduction that would be expected from the increase in

**Table 2.  2000 A.C.E. Coverage Estimates for Major Demographic Groups**

| Characteristic | Net Undercount (%) | Coverage Correction Factor | Data-Defined Rate | Correct Enumeration Rate | Inverse of Match Rate |
|---|---|---|---|---|---|
| Total | 1.18 | 1.0119 | .9707 | .9528 | 1.0918 |
| **Race/Origin Domain** | | | | | |
| Non-Hispanic White | 0.67 | 1.0068 | .9770 | .9590 | 1.0735 |
| Non-Hispanic Black | 2.17 | 1.0221 | .9565 | .9273 | 1.1504 |
| Hispanic | 2.85 | 1.0294 | .9521 | .9446 | 1.1431 |
| Hawaiian or Pacific Isl | 4.60 | 1.0483 | .9541 | .9305 | 1.1777 |
| Non-Hispanic Asian | 0.96 | 1.0097 | .9649 | .9457 | 1.1058 |
| AI On Reservation | 4.74 | 1.0498 | .9413 | .9581 | 1.1629 |
| AI Off Reservation | 3.28 | 1.0339 | .9624 | .9397 | 1.1382 |
| **Tenure** | | | | | |
| Owner | 0.44 | 1.0045 | .9761 | .9641 | 1.0661 |
| Non-Owner | 2.75 | 1.0283 | .9590 | .9269 | 1.1551 |
| **Age/Sex** | | | | | |
| 0-17 | 1.54 | 1.0157 | .9600 | .9594 | 1.1008 |
| 18-29 Male | 3.77 | 1.0391 | .9635 | .9290 | 1.1562 |
| 18-29 Female | 2.23 | 1.0228 | .9654 | .9362 | 1.1293 |
| 30-49 Male | 1.86 | 1.0190 | .9747 | .9522 | 1.0960 |
| 30-49 Female | 0.96 | 1.0097 | .9763 | .9600 | 1.0762 |
| 50+ Male | -0.25 | 0.9975 | .9796 | .9535 | 1.0673 |
| 50+ Female | -0.79 | 0.9922 | .9792 | .9552 | 1.0604 |

Net undercount is for household population.
A negative net undercount denotes a net overcount.

sample size alone. For the major demographic groups in this paper, the sampling variances were also generally smaller than expected.

## B. DSE Components

For each Census 2000 race/Hispanic origin domain, tenure category, and age/sex group, Table 2 summarizes the following estimates:

- percent net undercount
- coverage correction factor
- percent of data-defined people
- correct enumeration rate
- inverse of match rate

The coverage correction factor is obtained by multiplying the percent data-defined by the correct enumeration rate and the inverse of the match rate. Any differences are due to rounding.

Table 2 shows that 97.07 percent of all people in the census were data-defined. The Non-Hispanic White or "Some other race" domain had the highest percentage of data-defined people at 97.7 percent. The American Indian On Reservation domain had the lowest percentage of data-defined people at 94.13 percent.

Owners had a higher proportion of data-defined persons than Non-Owners. Children had the lowest data-defined percentage (96 percent) of the seven age/sex groups with 50+ Males having the highest proportion of data-defined persons (97.96 percent). The data-defined rates are variable within the race/Hispanic origin domains and the age/sex groups, but note that some of the variability may be due to small sample sizes.

The correct enumeration rate is a weighted estimate of the number of correctly enumerated people in the E Sample. The overall correct enumeration rate for the U.S. was 95.28 percent. The Non-Hispanic White or "Some other race" domain had a higher correct enumeration rate (95.9 percent) than any other race/Hispanic origin domain. The lowest correct enumeration rate was for the Non-Hispanic Black domain at 92.73 percent. As expected, Owners had a higher correct enumeration rate than Non-Owners. Females who are 30-49 years old had the highest correct enumeration rate (96 percent), closely followed by children (95.94 percent) and 50+ Females (95.52 percent).

The age/sex category with the lowest correct enumeration rate was 18-29 year-old Males (92.9 percent).

The match rate is the ratio of P Sample matches to persons in the P Sample. The inverse of the match rate estimates the adjustment for persons found in the P Sample but not in the census. The overall match rate was 91.59 percent. The lowest match rate of the seven race/Hispanic origin domains was 84.91 percent, corresponding to the Native Hawaiian or Other Pacific Islander domain. The Non-Hispanic White or "Some other race" domain had the highest match rate at 93.15 percent. Owners had a higher match rate than Non-Owners. The match rate for Owners was 93.8 percent while Non-Owners had a match rate of 86.57 percent. The 50+ Female and Male groups had the highest match rates (94.3 and 93.69 percent, respectively). The 18-29 year-old Male and Female groups had the lowest match rates of 86.49 and 88.55 percent, respectively.

## VI. Conclusions

Dual system estimation shows that Census 2000 undercounted the national household population and differentially undercounted population subgroups. Relative to the 1990 census, Census 2000 showed measured improvement in the overall percent net undercount and the differential undercounts of certain population groups.

## VII. References

Bryant, B. E. et al. (1992). "Assessment of Accuracy of Adjusted Versus Unadjusted 1990 Census Base for Use in Intercensal Estimates: Recommendation," Report of the Committee on Adjustment of Postcensal Estimates, 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* (See http://www.census.gov/dmd/www/pdf/Fr9.pdf)

Haines, D. (2000). "Accuracy and Coverage Evaluation Survey: Final Post-stratification Plan for Dual System Estimation," DSSD Census 2000 Procedures and Operations Memorandum Series # Q-24.

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.

Haines, D.E. and Hill, J.M. (1998). A Method for Evaluating Alternative Raking Control Variables. American Statistical Association Proceedings of the Survey Research Methods Section, 647-652.

Kim, J. K., Navarro, A. and Fuller, W. A. (2000). Variance Estimation for 2000 Census Coverage Estimates. American Statistical Association Proceedings of the Survey Research Methods Section, 515-520.

Navarro, A. and Sands, R. D. (2001). 2000 Census A.C.E. Variance Estimates. American Statistical Association Proceedings of the Survey Research Methods Section.

Thompson, J. H. (1992). "CAPE Processing Results," U.S. Census Bureau Memorandum, Washington, D.C.