Census 2000 Evaluation N.17
September 5, 2002

Targeted Extended Search Analysis

FINAL REPORT

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

Glenn Wolfgang,
Phawn Stallone, and
Tamara Adams

Decennial Statistical
Studies Division

U S C E N S U S B U R E A U

Helping You Make Informed Decisions

CONTENTS

LIST OF TABLES

EXECUTIVE SUMMARY

The Targeted Extended Search extended the search for persons who, due to geocoding error, would otherwise have been coded as missed or erroneously enumerated. Geocoding error is the incorrect assignment of a housing unit to a census block and cluster identification code (geocode). The usual area of search was the sample block cluster. Targeted Extended Search did extend search to blocks surrounding the sample cluster. Strategies for targeting clusters and addresses were developed to enhance efficiency in field and processing workloads. Targeted Extended Search aimed to reduce dual system estimate variances inflated by census geocoding errors and provide robustness against the Accuracy and Coverage Evaluation geocoding errors. Balance between the number of matches found in surrounding blocks and the number of correct enumerations found in surrounding blocks was important to Targeted Extended Search.

The goals of this Targeted Extended Search analysis were to report overall results and to identify characteristics that may be related to matches and correct enumerations found in surrounding blocks due to geocoding error. Understanding Targeted Extended Search results is essential to evaluating coverage measurement procedures in Census 2000. Any unexpected results could raise questions regarding the quality of the Accuracy and Coverage Evaluation methods and operations. Results regarding the impact on dual system estimation variances and the balance in Targeted Extended Search results were reported by other evaluations.

Specifically, we analyzed percentages of matches and correct enumerations found in surrounding blocks by categories of age, sex, tenure, race, Hispanic origin, region, and other operational variables to measure their effect as predictors of those percentages. Some computational simplifications increased the versatility and efficiency of this research. Statistics in this study were computed using nonmovers and outmovers without adjustment for the number of inmovers, unlike official statistics which did adjust for inmovers. Variances were estimated without fully accounting for all aspects of the sampling and estimation, such as missing data imputation variance estimation.

Did targeting clusters and addresses for extended search enhance efficiency in field and processing workloads?

Yes, targeting kept the 2000 search operations about half the size they would have been if 1990 search procedures would have been used.

- While about six percent of weighted P-sample persons were identified as Targeted Extended Search persons, about twelve percent, those not matched in sample clusters, would have needed extended search if there were no targeting.

- While about 3.2 percent of weighted E-sample persons were identified as Targeted Extended Search persons, about 5.1 percent, those not confirmed as correct enumerations in sample clusters, would have needed extended search processing if there were no targeting.

Did any results point to particular variables as predictors of percentages of matches or correct enumerations in surrounding blocks?

Yes, but the results provided few clues of how variables related to geocoding errors.

- Geocoding error coincided with coverage error in some analyses. This pattern of results recurred in a small number of the analyses: the analysis variable subgroups with a high overall percent not matched also had a high percent of matches in surrounding blocks. This pattern was found for persons who do not own their homes, persons in subsampled clusters, single family homes, and proxy-response households. A similar pattern was found in only one E-sample analysis: for home-owners, the overall percent of erroneous enumerations correlated with the percent of correct enumerations in surrounding blocks. This pattern failed to appear in most P-sample and E-sample analyses.

- Analyses of variables named Metropolitan Statistical Area / Type of Enumeration Area, Region, Subsampling Involvement, and Type of Structure at Basic Address gave partial support to the view that differences in geocoding error would be found between levels of geography or other characteristics that define the nature of the whole cluster. That view did not explain all the differences found.

- There were few indications of any other relationship between post-stratification or other operational variables and the surrounding block statistics. None of the variables tested provided an unqualified means to better understand or manage geocoding error.

What were the implications of these results?

Targeted Extended Search improved operational efficiency relative to the 1990 extended search operation. Evidence in summary data of an overall lack of balance between numbers of surrounding block matches and surrounding block correct enumerations prompted further data collection and evaluation reported by Adams and Liu (2001). Also, another study of the reduction of dual system estimate variances by the Targeted Extended Search was reported by Navarro and Olson (2001). Explorations of the relationships between other variables and Targeted Extended Search cases provided no cause for concern or insight for better managing geocoding error.

1. BACKGROUND

This report summarizes Targeted Extended Search (TES) procedures and some analyses results concerning TES effects on geocoding errors in the Accuracy and Coverage Evaluation (A.C.E.).

1.1 What was the role of the TES in the Census 2000 A.C.E.?

A goal of Census 2000 was to enumerate each person once at their residence on Census Day, April 1, 2000. The A.C.E. estimated the number of persons missed or erroneously enumerated in the census. One step in the estimation process involved matching persons enumerated in the A.C.E. clusters to persons enumerated in the census.

The TES extended the search beyond the cluster for missed enumerations or erroneous enumerations attributable to geocoding error. Geocoding error is the incorrect assignment of a housing unit to a census block and cluster identification code (geocode). The usual area of search was the sample block cluster. TES extended the search operation to blocks surrounding the sample cluster. TES was targeted at clusters and blocks most likely to benefit in efforts to reduce the effects of geocoding error in estimation (Childers, 2001). Surrounding block search operations were developed in past census coverage evaluations (Childers, 1999). TES reduced dual system estimate variances inflated by census geocoding errors and provided robustness to A.C.E. geocoding errors (Navarro and Olson, 2001).

1.2 What kinds of geocoding errors were targeted in the TES?

Two types of census geocoding errors and one type of A.C.E. geocoding error occurred:

- **Census errors of inclusion** occurred when a housing unit physically located outside a sample cluster erroneously had an in-cluster block number on census records. In this situation, a field search for the address limited within the cluster would have failed to confirm that it was correctly enumerated by the census. As a result of the TES field followup finding the address in the search area, the people enumerated at that address who would have been census errors of inclusion and coded as erroneous enumerations were recoded as duplicates or correct enumerations.

- **Census errors of exclusion** occurred when a housing unit physically located within a sample cluster erroneously had a not-in-cluster block number on the census record. In this situation, persons listed in the A.C.E. may not have been matched to census persons enumerated at that address. When the search was extended to census records in the surrounding blocks and clerks found matching persons, P-sample persons who would have been census errors of exclusion and coded as nonmatches were instead recoded as matches.

- **A.C.E. geocoding errors** occurred when a housing unit physically located outside a sample cluster had an A.C.E. address record with an in-sample block cluster number. In this situation, as with census errors of exclusion, persons listed in the A.C.E. may not have been matched to census persons enumerated at that address. When the search was extended to

census records in the surrounding blocks and clerks found matching persons, P-sample persons who would have been A.C.E. geocoding errors and coded as nonmatches were instead recoded as matches.

1.3 What aspects of A.C.E. operations were relevant to the TES?

The A.C.E. sample consisted of block clusters; large block clusters were subsampled (a separate and subsequent operation to the reduction in the number of small and medium block clusters). There were 11,303 A.C.E. clusters in the 50 states and District of Columbia. The addresses and people enumerated by the census in the A.C.E. clusters comprised the E sample, which was used to measure errors in the census data. The A.C.E. also independently listed the addresses and persons in the sample areas. Those addresses and persons comprised the P sample, which was used to measure what the census missed.

Data collection and processing began with the housing unit phase, in which addresses in the sample were listed and confirmed. Processing staff matched addresses independently listed by the A.C.E. to census housing unit addresses in the Decennial Master Address File. An address found in both sources was called a match; an address not found in both was a nonmatch. Staff conducted a housing unit field followup to confirm the existence of nonmatched housing units and to resolve other incomplete housing unit information.

In the person phase, A.C.E. field staff conducted independent interviews to obtain data on Census Day residents at the addresses listed in the housing unit phase of the A.C.E. The people listed in the A.C.E. housing units were P-sample persons. The people enumerated by the census in the A.C.E. sample areas were E-sample persons. Staff matched P-sample people to census enumerations. An operation called person followup collected any additional information in the field needed to resolve match status, residency on Census Day, correct enumeration status, or to confirm geocoding for specific cases.

1.4 How did the TES contrast to the 1990 Surrounding Block Search?

In 2000, the TES involved searching outside the sample cluster for persons or housing units that were likely affected by geocoding error. The area in which the A.C.E. looked for matches and correct enumerations was called the search area. The block cluster was the search area for non-TES cases. For TES cases, the search area was extended to the first ring of surrounding blocks, which included any block touching the A.C.E. cluster at any point, even if only at a corner point. The extended search was targeted to selected clusters and selected households in those clusters that showed potential for geocoding error, as measured by A.C.E. housing unit phase results and confirmed with person matching whole-household nonmatch results.

The TES differed from the 1990 surrounding block search primarily in its focus on geocoding error. The 1990 procedures extended the search area for matches, correct enumerations, or duplicates for all clusters. In contrast, the TES targeted clusters and specific households where the effort would be most beneficial for handling geocoding error. Another difference was that the search area was limited to one ring of surrounding blocks for the TES. In 1990, a second ring

of surrounding blocks (comprised of blocks not touching the cluster but not more than one block away) was searched for update/leave type of census enumeration areas, and the entire address register area was searched for list/enumerate clusters. The TES was designed to improve the operational efficiency and the quality of the extended search over 1990 procedures (Childers, 2001).

1.5 What were the effects of TES on A.C.E. estimates?

The TES identified matches and correct enumerations within the search area in housing units with geocoding errors. Without the TES, the number of both would have been lower. A few duplicates were also identified but remained erroneous enumerations and did not affect the number of correct enumerations. The dual system estimate (DSE) formula shows the role of matches (M) and correct enumerations (CE), expressed as the proportions (M/P and CE/E) of their respective samples, along with the number of data-defined census persons excluding late census adds and whole-person imputations (DD) (Davis, 2001). :

$$DSE = DD * (CE/E) / (M/P)).$$

One can see from the DSE formula that if the number of matches and correct enumerations increased about the same amount by the extended search, given that E and P samples had about the same total size, the expected value of the DSE will not be affected. Because A.C.E. sampling was random, any given case of census geocoding error affecting the sample should be as likely included erroneously as excluded erroneously. As long as the search areas for P-sample and E-sample cases are kept the same, census errors of inclusion and census errors of exclusion should be equal and balance each other in the sense of changing matches and correct enumerations at the same rate. Mulry and Spencer (1991) discussed balancing.

The A.C.E. geocoding errors, on the other hand, were not balanced with anything else. Like extended searches in previous censuses, the TES was designed to provide robustness against A.C.E. geocoding error. Navarro and Olson (2001) and Adams and Liu (2001) evaluated and reported on the potential lack of balance in the TES. Navarro and Olson also reported preliminary analysis of the TES impact on variances.

1.6 What other reports provided background to the TES?

Other reports provided greater detail on the A.C.E. and prior census coverage evaluations. Hogan (1993) reported on both analyses and procedures for the 1990 census. Hogan (2000) described application of theory in the A.C.E. Childers (2001) described the A.C.E. design. Adams, Barrett, and Byrne (2001) summarized procedures for A.C.E. operations.

This report's results were related to other research that divided the P and E samples into groups on the basis of levels of important variables and tested for differences in percentages not matched or percentages of erroneous enumerations. P-sample percentages not matched were reported by Wolfgang, Adams, Davis, Liu, and Stallone (2001). Feldpausch (2001) investigated E-sample percentages of erroneous enumerations in Census 2000. Jones (2002) compared the percentages of geocoding errors identified in housing unit operations prior to A.C.E. person interviewing.

2. METHODS

2.1 How was the Targeted Extended Search conducted?

Extended search procedures involved both clerical matching of data records and field followup visits. The TES field followup was conducted about the same time as A.C.E. person interviewing. In clusters selected for the TES, the field staff canvassed the cluster or surrounding blocks to locate census housing units identified as potential geocoding errors. If the housing unit was found in the search area, the data of persons enumerated at that unit were reviewed to determine if those persons were enumerated correctly or duplicated in the census.

The potential for geocoding error was determined in the housing unit matching and followup preceding person interviewing. Nonmatched E-sample addresses confirmed by field followup to exist as a housing unit outside the sample cluster point to census errors of inclusion. Nonmatched P-sample addresses suggest census errors of exclusion or A.C.E. geocoding errors. A cluster's sum of these nonmatched addresses in both E and P samples was the measure of the potential geocoding error used for targeting clusters for the TES.

The targeting of potential geocoding error cases was designed at both cluster and household levels (Childers, 2001):

- **Cluster Targeting and Sample Selection** – Targeting clusters reduced the number of clusters selected for extended search by nearly 80 percent. First, TES selection included with certainty 62 clusters for which housing unit matching was too delayed to ascertain the geocoding status of the census units in those clusters. List/enumerate clusters, for which census data were not available in time for TES field followup, were excluded from the TES. From the remaining clusters, 1,088 with the highest weighted and unweighted measures of geocoding error potential were selected for the TES with certainty. Another 1,089 clusters were selected by sampling from those with a non-zero count of potential geocoding error.

- **Address Targeting** – Rather than processing every address in TES clusters, the operation targeted households with characteristics of geocoding error. Specifically, the TES targeted addresses as follows:

    - P sample – During person matching, matching staff searched the census data of surrounding blocks only for whole-household nonmatches (that is, residents of households in which all persons were nonmatched) in any P-sample nonmatched housing unit (including units for which matching census housing units were deleted after the matching operation was done). Matches resulting from these searches were attributable to census errors of exclusion or A.C.E. geocoding errors. We also limited our surrounding block search in urban areas to the block in which a matching census address was found. In clusters with one or more non-city-style address, we searched in all surrounding blocks.

- E sample – During the time of person interviewing, a TES field followup confirmed whether E-sample addresses coded as potential census geocoding error were located in the search area. Only whole-household nonmatches at addresses confirmed to be in the search area were attributed to census errors of inclusion and recoded as correct enumerations. A duplicate search for any people coded outside the cluster was limited to the block in which the housing unit was located in TES field followup.

2.2 What other operations or characteristics were important to TES ?

Subsampling and imputation were important A.C.E. estimation operations. If there were eighty or more housing units in a block, a block segment was subsampled within the block. Values were imputed, if missing, for variables used in post-stratification, namely tenure, age, sex, race, and Hispanic origin.

Subsampling, imputation, and characteristics of housing units or persons (like type of respondent, type of address, mover status, household size, or kinship within the household) were not intended to interact with TES results. But it could be instructive to find a relationship between them and percentages of matches or correct enumerations found by targeted extended search in surrounding blocks. That is the reason for many of the analyses below.

If match status remained unresolved, a match probability was assigned. If residence status remained unresolved, a residence probability was assigned. For households not successfully interviewed, a non-interview adjustment was applied. Sampling weights, including weights for large block subsampling and TES selection, match probabilities, residence probabilities, and non-interview adjustments were applied in all analyses.

2.3 How were results analyzed?

In addition to reporting some general results, this study analyzed differences among percentages of surrounding block matches and percentages of surrounding block correct enumerations. Those statistics were computed within subgroups of the sample defined by levels of descriptive variables. Specifically, surrounding block matches were matches assigned only due to TES operations and the denominator for the percent of surrounding block matches was the number of P-sample persons in the subgroup. Surrounding block correct enumerations similarly were assigned only in TES operations. The denominator for the percent of surrounding block correct enumerations was the number of E-sample persons in the subgroup. Full weighting, based on both the general and the TES sampling, of these numbers was used. Only the use of inmover data was omitted. Groups with high percentages of surrounding block matches or surrounding block correct enumerations would provide insight on conditions associated with geocoding error.

The percentages of matches or correct enumerations in surrounding blocks for different sample subgroups were compared using stratified jackknife variance estimation and pair-wise t values generated by VPLX (Fay, 1990). Variances were confirmed by an alternate program.

Statistical significance for these t values was determined using the Bonferroni multiple comparison of means technique. It controlled 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. LIMITATIONS

There were certain limitations in the results presented in this paper. Several were computational shortcuts that permitted the efficiency and versatility needed to conduct a wide range of analyses.

- Inmover data were not used as in official dual system estimates. 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 poststratum variables. In these analyses, we were interested in some non-post-stratum variables and used the simpler methodology. See Haines (2001b) for a description of the conditions and methods for using inmover data in official estimates. Match rates computed in the production of official statistics (Davis, 2001) typically were 0.3 percent lower than those generated with computational methods as in this report (Wolfgang, Adams, Davis, Liu, and Stallone, 2001). So differences between pairs of rates were much the same with one computation as the other. If the statistics computed in this study follow that pattern, differences tested in this analysis should be about the same with or without using the inmover data.

- Match statistics excluding surrounding block matches were not precisely the same as match results if no TES were done, because followup and estimation might affect the residence status or weighting of the persons involved.

- Standard error computations in these analyses were simplified and did not take into account all aspects of the sampling and estimation, such as missing data imputation variance estimation.

4. RESULTS

A few overall results showed the effects of geocoding error removal from the A.C.E. by the TES, increasing matches and correct enumerations in the A.C.E. Other analyses compare the percentages of matches and correct enumerations found in surrounding blocks for different sample subgroups. They show which, if any, variables are related to the impact of geocoding error.

4.1 What were the weighted numbers of matches, correct enumerations, and duplicates in sample clusters and in surrounding blocks?

Table 4.1.1 shows that due to targeting, which focused efforts on potential geocoding error cases or a representative sample of those cases, only 6.0 percent of weighted P-sample persons and 3.2 percent of weighted E-sample persons were involved in TES processing. Certain E-sample cases in relisted clusters had been given TES status before being found in the cluster; they were reassigned as not TES persons (Beaghen, 2001).

The size of the TES in the P sample was much larger than in the E sample because selection of cases with potential for geocoding error was more precise in the E sample. More of the erroneous enumerations selected for the TES, relative to nonmatches selected for the TES, were recoded due to geocoding error.

Table 4.1.1. Number of Persons in the TES in P and E samples

	Weighted P sample	Weighted E sample
Total Sample	258,547,382	264,578,862
non TES	243,077,600	256,034,032
TES	15,469,782	8,544,830
TES(%)	6.0%	3.2%

Table 4.1.2 shows that the surrounding block matches made up about 3.9 percent of the P sample, using nonmover and outmover data without any inmover data information. Navarro and Olson (2001), using inmover data as in official computations, report 3.8 percent. These TES cases matched in surrounding blocks represent the sum of geocoding error due to census errors of exclusion and A.C.E. geocoding errors. The two types of error cannot be distinguished using information available to this study.

Table 4.1.2. Matches in Surrounding Blocks

	Weighted Number	Weighted Percent
Total P sample	258,547,382	100.0
Matches:		
Total	237,401,214	91.8
In Sample Clusters	227,399,141	87.9
In Surrounding Blocks	10,002,073	3.9

Table 4.1.3 shows that surrounding block correct enumerations made up about 2.9 percent of the E sample. These TES cases coded correctly enumerated in surrounding blocks represented geocoding error due to census errors of inclusion.

Table 4.1.3. Correct Enumerations in Surrounding Blocks

	Weighted Number	Weighted Percent
Total E sample	264,578,862	100.0
Correct Enumerations:		
Total	252,096,238	95.3
In Sample Clusters	244,387,951	92.4
In Surrounding Blocks	7,708,287	2.9

The difference in the percents of matches and correct enumerations in surrounding blocks was explained as due largely to A.C.E. geocoding errors (Adams and Liu, 2001).

If 1990 extended search procedures had been used, the size of the extended search operation would have been about double that of the TES. For the P-sample, all persons not matched in the sample clusters, about 12 percent (See Matches In Sample Clusters in Table 4.1.2) would have needed extended search if there were no targeting. For the E sample, 7.6 percent of persons were not confirmed as correct enumerations in the sample clusters (See Table 4.1.3). Excluding the 0.7 percent which were duplicates (See Table 4.1.4) and the 1.8 percent which would not be processed due to insufficient information (Zelenak, 2001), 5.1 percent would have needed extended search processing if there were no targeting.

The contrast between the last two rows of Table 4.1.4 shows that most of the duplicates found between E-sample persons and other census enumerations were found within the sample clusters; that is, both records had sample cluster identifications. TES found a small number of census enumerations in surrounding blocks that duplicated E-sample person records. Census duplicate delete and reinstated cases were not available in the E-sample data processed for the A.C.E.

Table 4.1.4. Duplicates in Surrounding Blocks

	Weighted Number	Weighted Percent
Total E sample	264,578,862	100.00
Duplicates:		
Total	1,852,499	0.70
In Sample Clusters	1,759,313	0.66
In Surrounding Blocks	93,186	0.04

4.2 What variables related to percent of matches in surrounding blocks or the percent of correct enumerations in surrounding blocks?

This section provides results for relating other characteristics of clusters, households, or persons to percent of matches or correct enumerations in surrounding blocks; those statistics are measures of the TES impact on the A.C.E. Results are presented in tables displaying in each row: variable level name, percent of matches or correct enumerations in surrounding blocks (under the column heading "Percent"), the rank of that level's percent not matched from lowest to highest ("Rank"), a list of the ranks for other variable levels with which a significant difference was found ("Differs from"), the stratified jackknife standard error ("s.e."), and the weighted percent of persons comprising that level's subgroup ("n(percent)"). The denominator of percents in the "n" column is the weighted total number of persons in the sample, as found in Table 4.1.1. Percents may not sum precisely to 100 due to rounding error. Criterion t-values (e.g., "$|t| > 1.65$", noted below each table) varied, as described above, with the number of comparisons being made in the family of tests.

Important variables were grouped into two categories:

- Variables used in defining post-strata for dual system estimation
- Other variables relevant to sampling or estimation of TES

4.2.1 What post-stratification variables related to the percent of matches in surrounding blocks or the percent of correct enumerations in surrounding blocks?

Variables used to form post-strata in dual system estimation (Haines, 2001b) were of primary interest. They were analyzed here using the levels as defined for post-stratification. Levels of Metropolitan Statistical Area / Type of Enumeration Area (MSA/TEA), Return Rate Indicator, and Region were used in some but not all post-stratum group definitions. All P-sample persons were included in these analyses, regardless of whether their post-stratum was affected by the variable.

Analyses of Tenure yielded a pattern suggesting a relationship of surrounding blocks results to overall within-cluster results. Wolfgang, Adams, Davis, Liu, and Stallone (2001) report that those who did not own their homes had a higher overall percent not matched in the whole P sample (13.1 as compared to 6.1 for owners). At the same time, that group had a higher

percent of matches in surrounding blocks (See Table 4.2.1). In other words, more matches in surrounding blocks were found for the group where there were more nonmatches in general. Without TES, matches in surrounding blocks would have been additional nonmatches.

Similarly, Feldpausch (2001) reports that E-sample non-owners had a higher percent of erroneous enumerations. They also had a higher percent of correct enumerations in surrounding blocks (See Table 4.2.2). More correct enumerations in surrounding blocks were found for the group where there were more erroneous enumerations in general. Without TES, correct enumerations in surrounding blocks would have been additional erroneous enumerations.

This pattern of results suggests that geocoding errors might coincide with coverage errors, specifically, nonmatches and erroneous enumerations. This hypothetical relationship does not imply one type of error causes the other. A geocoding error case cannot be a coverage error case; no single case can contribute to both a higher percent of overall nonmatches and a higher percent of matches in surrounding blocks. Similarly, no single case can contribute to both a higher percent of overall erroneous enumerations and a higher percent of correct enumerations in surrounding blocks. However, Tenure was the only variable that yielded that relationship across all groups in both E sample and P sample. The pattern was partially observed in analyses discussed below, but was not observed consistently enough to be considered generally applicable.

Indeed, only MSA/TEA, among other post-strata variables, yielded more than a few scattered significantly different test results, and its pattern suggests another hypothetical rule: that variable groups which differ in geographical features may yield differences in these statistics related to geocoding error.

Table 4.2.1. Percent of Matches in Surrounding Blocks, by Tenure

Tenure	Percent	Rank	Differs from	s.e.	n (percent)
Owner	3.3	1	2	0.2	69.8
Non-owner	5.1	2	1	0.6	30.2

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

Table 4.2.2. Percent of Correct Enumerations in Surrounding Blocks, by Tenure

Tenure	Percent	Rank	Differs from	s.e.	n (percent)
Owner	2.7	1	2	0.2	69.7
Non-owner	3.4	2	1	0.3	30.3

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

Metropolitan statistical areas delineate cities for statistical purposes. Type of Enumeration Area describes the method of data collection adopted for an area. About 82 percent of the sample persons live in Mailout/Mailback type of enumeration areas, where Census 2000 forms were

mailed to their addresses with directions to return responses by mail. Other TEAs involve Census 2000 staff bringing forms to housing units in the area, usually updating address listings and leaving the forms for mailback, but sometimes listing addresses for the first time or collecting enumerations on the spot. Metropolitan Statistical Area size and Type of Enumeration Area were combined into one variable (MSA/TEA) used for post-stratification.

All levels of Mailout/Mailback type of enumeration area stood out with higher percentages of matches in surrounding blocks and of correctly enumerated in surrounding blocks than the level combining all other types of enumeration area (See Tables 4.2.3 and 4.2.4). Those results do not parallel those for nonmatches and erroneous enumerations within sample blocks as for Tenure, but instead seemed to reflect differences in geography or postal service organization associated with enumeration by Mailout/Mailback. Jones (2002) similarly found Mailout/Mailback areas had a higher percentage of geocoding errors than areas where addresses were updated and forms left by Census Bureau staff.

Table 4.2.3. Percent of Matches in Surrounding Blocks, by Size of Metropolitan Statistical Area and Type of Enumeration Area

MSA/TEA	Percent	Rank	Differs from	s.e.	n (percent)
Large MSA, Mailout/Mailback	4.4	3	1	0.5	30.4
Medium MSA, Mailout/Mailback	4.4	4	1	0.5	31.3
Small MSA & Non-MSA Mailout/Mailback	4.0	2	1	0.4	20.2
All Other TEAs	1.9	1	all	0.3	18.1

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

Table 4.2.4. Percent of Correct Enumerations in Surrounding Blocks, by Size of Metropolitan Statistical Area and Type of Enumeration Area

MSA/TEA	Percent	Rank	Differs from	s.e.	n (percent)
Large MSA, Mailout/Mailback	2.8	2	1	0.3	30.2
Medium MSA, Mailout/Mailback	3.9	4	1	0.4	31.5
Small MSA & Non-MSA Mailout/Mailback	3.5	3	1	0.3	20.4
All Other TEAs	0.7	1	all	0.1	17.9

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

Few of the tests in the eight other post-stratification variable tables were significant. Only three of the tables, all surrounding block correct enumeration analyses, had any significant differences,

and in each of them there was but one group that differed from some but not all other groups: Females aged 18-29 (3.5; see Table 4.2.6) , Native Hawaiian or Pacific Islander (1.1; see Table 4.2.8), and the Midwest region (2.2; see Table 4.2.12). Because those differences were few, without any evident pattern, and without replication in other times and places, those results could be explained as Type I error and not worth much attention.

Age/Sex results (See Tables 4.2.5 and 4.2.6) illustrate the difficulty of meaningful interpretation given a weak, inconsistent pattern of results. None of the percentages of matches in surrounding blocks differed significantly. Only three of twenty-one tests of the Age/Sex percents of correct enumerations were significantly different. The pattern in all the Age/Sex or Race and Hispanic Origin Domain results (See Tables 4.2.7 and 4.2.8) could be explained as random variation. See the discussion of Tables 4.2.11 and 4.2.12 for interpretation of the region analyses.

Table 4.2.5. Percent of Matches in Surrounding Blocks, by Age and Sex

Age/Sex	Percent	Rank	Differs from	s.e.	n (percent)
0-17	3.9	5	none	0.3	26.2
18-29 Male	3.8	2	none	0.3	7.5
18-29 Female	4.1	7	none	0.3	7.7
30-49 Male	3.8	3	none	0.2	15.2
30-49 Female	4.0	6	none	0.3	16.2
50+ Male	3.7	1	none	0.2	12.3
50+ Female	3.8	4	none	0.3	15.0

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

Table 4.2.6. Percent of Correct Enumerations in Surrounding Blocks, by Age and Sex

Age/Sex	Percent	Rank	Differs from	s.e.	n (percent)
0-17	2.8	2	7	0.2	25.7
18-29 Male	3.1	6	none	0.3	7.8
18-29 Female	3.5	7	2,4,5	0.3	7.8
30-49 Male	2.9	4	7	0.2	15.3
30-49 Female	2.9	5	7	0.2	15.8
50+ Male	2.8	1	none	0.2	12.6
50+ Female	2.9	3	none	0.2	15.0

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

Table 4.2.7. Percent of Matches in Surrounding Blocks, by Race and Hispanic Origin

Race and Hispanic Origin Domain	Percent	Rank	Differs from	s.e.	n (percent)
American Indian on Reservation	2.3	1	none	0.8	0.2
American Indian off Reservation	3.9	3	none	1.3	0.5
Hispanic	4.2	5	none	0.6	12.3
Non-Hispanic Black	4.5	6	none	0.6	11.4
Native Hawaiian or Pacific Islander	5.3	7	none	1.9	0.2
Non-Hispanic Asian	4.2	4	none	0.8	3.4
Non-Hispanic White	3.7	2	none	0.2	72.1

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

Table 4.2.8. Percent of Correct Enumerations in Surrounding Blocks, by Race and Hispanic Origin

Race and Hispanic Origin Domain	Percent	Rank	Differs from	s.e.	n (percent)
American Indian on Reservation	1.3	3	none	0.9	0.2
American Indian off Reservation	2.0	2	none	0.5	0.5
Hispanic	2.4	4	none	0.3	12.3
Non-Hispanic Black	3.1	7	1	0.5	11.8
Native Hawaiian or Pacific Islander	1.1	1	5,6,7	0.4	0.2
Non-Hispanic Asian	3.1	6	1	0.5	3.7
Non-Hispanic White	3.0	5	1	0.2	71.4

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

The tract-level census return rate, a sign of public cooperation, was the proportion of occupied housing units in a census tract that returned a Census 2000 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, 2001b). No differences were evident in the return rate analyses (Tables 4.2.9 and 4.2.10).

Table 4.2.9. Percent of Matches in Surrounding Blocks, by Return Rate Indicator

Return Rate Indicator	Percent	Rank	Differs from	s.e.	n (percent)
High	3.9	2	none	0.3	72.3
Low	3.6	1	none	0.4	23.5
Not Applicable	4.1	3	none	0.7	4.3

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

Table 4.2.10. Percent of Correct Enumerations in Surrounding Blocks, by Return Rate Indicator

Return Rate Indicator	Percent	Rank	Differs from	s.e.	n (percent)
High	2.9	3	none	0.2	72.0
Low	2.9	1	none	0.3	23.5
Not Applicable	2.8	2	none	0.4	4.5

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

In all the tests of surrounding block matches and correct enumerations for Region (See Tables 4.2.11 and 4.2.12), only one was significant. In the E sample, the percent of correct enumerations found in surrounding blocks was lower for the Midwest relative to only the South. The same single significant result was found in the analysis of E-sample geocoding error recorded for housing units (Jones, 2002). In contrast, analyses of general nonmatches and erroneous enumerations for Region showed the Midwest region was consistently lower than other regions in rates. In only one of six tests did the results for surrounding block matches and correct enumerations coincide with those of general nonmatches and erroneous enumerations. Those results were not consistent enough to support an interpretation of the type offered for Tenure. At the same time, the results only weakly supported an hypothesis that geocoding error related to differences in geography.

Table 4.2.11. Percent of Matches in Surrounding Blocks, by Region of the United States

Region	Percent	Rank	Differs from	s.e.	n (percent)
Northeast	3.6	2	none	0.5	19.0
Midwest	3.1	1	none	0.3	22.9
South	4.4	4	none	0.5	35.4
West	4.0	3	none	0.5	22.8

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

Table 4.2.12. Percent of Correct Enumerations in Surrounding Blocks, by Region of the United States

Region	Percent	Rank	Differs from	s.e.	n (percent)
Northeast	3.2	3	none	0.4	19.1
Midwest	2.2	1	4	0.3	22.8
South	3.3	4	1	0.3	35.6
West	2.8	2	none	0.3	22.5

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

4.2.2 What other variables related to the percent of matches in surrounding blocks or the percent of correct enumerations in surrounding blocks?

Other operational or characteristic variables were analyzed. A pattern of results similar to the Tenure pattern was found for matches in surrounding blocks (but not necessarily for correct enumerations in surrounding blocks) in three other analyses described below: Subsampling Involvement, Type of Structure at Basic Address, and Type of Respondent. A geographical hypothesis also might fit for the Subsampling Involvement analysis. Analyses of other variables (Imputation of Characteristics, Mover Status, Household Size, and Kinship to Reference Person) yielded no significant differences or less recognizable patterns of results.

Subsampling of housing units within clusters with a large number of housing units was done to reduce the intra-cluster correlation and to reduce the interviewing workloads. The interesting result in this analysis is that one table below shows more geocoding error for the large blocks with subsampling, while the other shows less for that group. Geography did seem to relate to geocoding error, even if in contradictory ways from P sample to E sample.

Subsampled clusters had a higher percent not matched in the overall P sample and, at the same time, a higher percent of matches in surrounding blocks (Table 4.2.13). The hypothesized relationship of surrounding blocks results to overall within-cluster results might hold for the P sample, but the pattern is broken in the E sample, where subsampled clusters had a lower percent of correct enumerations in surrounding blocks (Table 4.2.14) in contrast to the result for matches in surrounding blocks.

Table 4.2.13. Percent of Matches in Surrounding Blocks, by Subsampling Involvement

Subsampled or Not	Percent	Rank	Differs from	s.e.	n (percent)
Not Subsampled	3.4	1	2	0.2	63.5
Subsampled	4.8	2	1	0.6	36.5

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

Table 4.2.14. Percent of Correct Enumerations in Surrounding Blocks, by Subsampling Involvement

Subsampled or Not	Percent	Rank	Differs from	s.e.	n (percent)
Not Subsampled	3.3	2	2	0.2	64.1
Subsampled	2.3	1	1	0.3	35.9

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

Type of structure at a basic address was collected basically as single-family dwelling versus multi-unit building in the census. P-sample data permit additional groupings for mobile homes and living quarters in a special place as well as unclassified structures. Ignoring the small number of living quarters in a special place and unclassified structures, which had large standard errors, the single-family dwellings had the smallest percent of matches in surrounding blocks (Table 4.2.15) and the smallest percent not matched over the whole P sample. Single-family dwellings had a lower percent of correct enumerations in surrounding blocks (Table 4.2.16) while, in a related but not quite parallel E-sample analysis (Feldpausch, 2001), addresses with only one housing unit had the lowest percent of erroneous enumerations. Jones (2002) found fewest geocoding errors at addresses with fewer than ten housing units. The hypothesis that geocoding errors parallel general coverage errors fits the P sample and seems to fit the E sample. If you view housing unit structure as something like geography (areas dominated by large multi-unit buildings contrasted to those that are not), a geographical hypothesis also seems to fit.

Table 4.2.15. Percent of Matches in Surrounding Blocks, by Type of Structure at Basic Address

Structure	Percent	Rank	Differs from	s.e.	n (percent)
Single-Family Dwelling	2.9	1	3,4	0.2	75.7
Multi-Unit	6.7	3	1	1.0	18.9
Mobile Home	7.1	4	1	0.9	5.2
Living Quarters in a Special Place and Unclassified	3.8	2	none	1.8	0.2

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

Table 4.2.16. Percent of Correct Enumerations in Surrounding Blocks, by Type of Structure

Structure	Percent	Rank	Differs from	s.e.	n (percent)
Single-Family Dwelling	2.5	1	2	0.2	79.2
Multi-Unit	4.4	2	1	0.5	20.8

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

The P-sample analysis of Proxy yielded a pattern similar to that for Tenure: the highest rate of matches attributable to geocoding error (Table 4.2.17) occur for proxies, which had the highest percent of nonmatches in general. But the E sample did not have that pattern. Proxies did not have more correct enumerations due to geocoding error (Table 4.2.18), however, although they did have more erroneous enumerations in general.

Table 4.2.17. Percent of Matches in Surrounding Blocks, by Type of Respondent

Type of Respondent	Percent	Rank	Differs from	s.e.	n (percent)
Proxy	5.2	2	1	0.5	5.5
Not Proxy	3.8	1	2	0.2	94.5

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

Table 4.2.18. Percent of Correct Enumerations in Surrounding Blocks, by Type of Respondent

Type of Respondent	Percent	Rank	Differs from	s.e.	n (percent)
Proxy	2.8	1	none	0.5	2.7
Not Proxy	2.9	2	none	0.2	97.3

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

There was no significant difference in surrounding block matches between those who did and those who did not move between Census Day and the A.C.E. interview (Table 4.2.19). Mover status is not relevant to the E sample; no corresponding analysis was done.

Table 4.2.19. Percent of Matches in Surrounding Blocks, by Mover Status

Person Mover Flag	Percent	Rank	Differs from	s.e.	n (percent)
Nonmover	3.9	1	none	0.2	96.6
Outmover	4.4	2	none	0.5	3.4

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

Age, sex, race, Hispanic origin, and tenure were sometimes imputed for the A.C.E. Where information was complete enough to require no imputation of post-stratum characteristics, the overall percent not matched and percent of erroneous enumerations were lower. The percent of matches in surrounding blocks did not differ (Table 4.2.20) and the percent of correct enumerations in surrounding blocks was significantly lower for the imputed group (Table 4.2.21), while the percent of erroneous enumerations in the overall sample was higher for those who had some imputation. The E-sample finding contradicts the hypothesis that geocoding errors parallel general coverage errors .

Table 4.2.20. Percent of Matches in Surrounding Blocks, by Imputation of Characteristics

Imputed or Not	Percent	Rank	Differs from	s.e.	n (percent)
Not Imputed	3.9	2	none	0.2	94.7
Imputed	3.4	1	none	0.4	5.3

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

Table 4.2.21. Percent of Correct Enumerations in Surrounding Blocks, by Imputation of Characteristics

Imputed or No	Percent	Rank	Differs from	s.e.	n (percent)
Not Imputed	3.0	2	1	0.2	87.0
Imputed	2.3	1	2	0.2	13.0

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

The Household Size and Kinship to Reference Person analyses had few significant results (Tables 4.2.22 to 4.2.25) or had another pattern that did not support any hypothesized explanation. For example, the Household Size percent of nonmatches in the overall sample was largest for the group of seven or more persons, while the percent of matches in surrounding blocks was smallest for that group (Table 4.2.22).

Table 4.2.22. Percent of Matches in Surrounding Blocks, by Household Size

Household Size	Percent	Rank	Differs from	s.e.	n (percent)
One person	4.6	3	1	0.4	10.5
2-6 persons	3.8	2	none	0.2	84.9
7 or more persons	2.9	1	3	0.5	4.7

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

Table 4.2.23. Percent of Correct Enumerations in Surrounding Blocks, by Household Size

Household Size	Percent	Rank	Differs from	s.e.	n (percent)
One person	3.7	3	2	0.3	10.1
2-6 persons	2.8	2	3	0.2	86.3
7 or more persons	2.6	1	none	0.5	3.6

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

Table 4.2.24. Percent of Matches in Surrounding Blocks, by Kinship to Reference Person

Kinship	Percent	Rank	Differs from	s.e.	n (percent)
Reference Person, Alone	4.6	5	none	0.5	10.1
Reference Person, Not Alone	3.9	4	2	0.2	28.3
Spouse	3.6	2	4	0.2	20.2
Parent / Child	3.9	3	none	0.3	30.4
Other relatives and nonrelatives	3.5	1	none	0.3	10.9

Note: Criterion for levels to differ was | t | > 2.568

Table 4.2.25. Percent of Correct Enumerations in Surrounding Blocks, by Kinship to Reference Person

Kinship	Percent	Rank	Differs from	s.e.	n (percent)
Reference Person, Alone	3.7	5	all	0.3	10.1
Reference Person, Not Alone	2.9	4	1,5	0.2	28.8
Spouse	2.8	1	4,5	0.2	20.3
Parent / Child	2.8	2	5	0.2	31.2
Other relatives and nonrelatives	2.9	3	5	0.2	9.5

Note: Criterion for levels to differ was $|t| > 2.568$

5. CONCLUSIONS

A few conclusions were drawn from this study:

- Targeting kept the 2000 search operations about half the size they would have been if 1990 search procedures would have been used.

- The TES found few duplicates in surrounding blocks. Census duplicate delete and reinstated cases were not available in the E-sample data.

- One pattern of results, suggesting a positive correlation between geocoding errors and coverage errors, recurred in a small number of the analyses. A variable subgroup with a high percent not matched also had a high percent of matches in surrounding blocks. In other words, evidence of more geocoding error was found for groups with more coverage error. This pattern was found for tenure, persons in subsampled clusters, single family homes, and proxy-response households. A similar pattern also was found in the E-sample analyses of tenure: the percent erroneous enumerations correlated with the percent of correct enumerations in surrounding blocks. Many other analyses failed to support and sometimes even contradicted the hypothesized relationship.

- Analyses of MSA/TEA, Region, Subsampling Involvement, and Type of Structure at Basic Address gave partial support to the view that differences in geography and other characteristics that define the nature of the whole block were related to geocoding error. But such variables do not explain all the differences found.

- There were few indications of a relationship between other post-stratification or operational variables and the surrounding block statistics. None of the variables tested provided an unqualified means to better understand or manage geocoding error.

- See Navarro and Olson (2001) and Adams and Liu (2001) for results concerning the balance between numbers of surrounding block matches and surrounding block correct enumerations.

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.

Adams, T. and Liu, X. (2001). "Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report Number 2: Evaluation of Lack of Balance and Geographic Errors Affecting Person Estimates" DSSD Census 2000 Procedures and Operations Memorandum Series T-13, U.S. Census Bureau, Washington, D.C.

Beaghen, M. (2001). "Accuracy and Coverage Evaluation: TES Balancing" DSSD Census 2000 Procedures and Operations Memorandum Series T-12, U.S. Census Bureau, Washington, D.C.

Childers, D. (1999). "Surrounding Block Search" unpublished internal document, U.S. Census Bureau, Washington, D.C.

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.

Fay , R. (1990). "VPLX: Variance Estimates for Complex Samples," Proceedings of the Section on Survey Research Methods, American Statistical Association, 266-271.

Feldpausch, R. (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, U.S. Census Bureau, Washington, D.C.

Haines, D. (2001a). "Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Dual System Estimation Output files," DSSD Census 2000 Procedures and Operations Memorandum Series Q-38, U.S. Census Bureau, Washington, D.C.

Haines, D. (2001b). "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.

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.

Jones, J. (2002), "The Analysis of Census Geocoding Error as Measured by the 2000 Accuracy and Coverage Evaluation", DSSD Census 2000 Procedures and Operations Memorandum Series U- [number to be assigned], U.S. Census Bureau, Washington, D.C.

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

Navarro, A. and Olson, D. (2001). "Accuracy and Coverage Evaluation: Effect of Targeted Extended Search," DSSD Census 2000 Procedures and Operations Memorandum Series B-18*, U.S. Census Bureau, Washington, D.C.

Wolfgang, G., Davis, P., and Stallone, P. (2001). "Accuracy and Coverage Evaluation Persons Not Matched in Census 2000," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.

Wolfgang, G., Stallone, P., and Adams, T. (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.

Wolfgang, G., Adams, T., Davis P., Liu, X., and Stallone, P. (2001). "Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report Number 18: P-sample Nonmatch Analysis" DSSD Census 2000 Procedures and Operations Memorandum Series T-15, U.S. Census Bureau, Washington, D.C.

Zelenak, M.F. (2001) . Accuracy and Coverage Evaluation: Control Tables for Missing Data and Estimation" DSSD Census 2000 Procedures and Operations Memorandum Series Q-57, U.S. Census Bureau, Washington, D.C.

APPENDIX

TECHNICAL DOCUMENTATION

A.1 Files and Data Sets Used

A.1.1 PFINUS.data:

Variables drawn from the P-sample Person Dual System Estimation final US data, as documented in Haines (2001a), and used in analyses for this report include:

- AGESEX Age/Sex Post-Stratification Variable
- AMTIMP Flag for any variables imputed
- DOMAIN A.C.E. Race/Hispanic Origin Domain, Post-Stratification Variable
- MOVERPER Person Mover Flag
- MPROB Match Probability (set to 0 for inmovers and nonresidents)
- MSATEA MSA/TEA Post-Stratification Variable
- POSSPSC Total Possible P-sample Persons in Household for Census Day
- PRXYIN Proxy/Nonproxy Respondent
- REGION Census Region of the United States, Post-Stratification Variable
- RELAT2 Recoded Relationship to Reference Person
- RPROB Residence Probability (set to 0 for inmovers)
- RRATEIND Return Rate Indicator, Post-Stratification Variable
- TENURE2 Recoded Tenure, Post-Stratification Variable
- TESFINWT P-sample Final TES-Adjusted Weight for Census Day
- TESPER TES Person Indicator
- TOBA Type of Basic Address

A.1.2 EFINUS.data:

Variables drawn from the E-sample Person Dual System Estimation final US data, as documented in Haines (2001a), and used in analyses for this report include:

- AGESEX Age/Sex Post-Stratification Variable
- AMTIMP Flag for any variables imputed
- DOMAIN A.C.E. Race/Hispanic Origin Domain, Post-Stratification Variable
- CEPROBF Probability of Correct Enumeration – Final
- MSATEA MSA/TEA Post-Stratification Variable
- NESAMP Total E-sample Persons in Household
- PRXYIN Proxy/Nonproxy Respondent
- REGION Census Region of the United States, Post-Stratification Variable
- RELAT2 Recoded Relationship to Reference Person
- RRATEIND Return Rate Indicator, Post-Stratification Variable
- STRCDE Structure Code
- TENURE2 Recoded Tenure, Post-Stratification Variable

- TESFINWT E-sample Final TES-Adjusted Weight for Census Day
- TESPER TES Person Indicator

A.1.3 SDF.US7:

The Sample Design File provided sample weighting and stratification information that was not included in the estimation files, including variables:

- STRATUM final sampling stratum ID – used in stratified jackknife variance estimation
- TESSELECT TES Selection Type
- WEIGHTC the cluster weight
- WEIGHTP the P-sample housing unit weight

A.1.4 PER.CUP:

This file was the census person record file generated from the Person Matching Review and Coding System. It was the source of:

- BFUMAT final PER BFU match code – used in case selection.

A.1.5 PROXY182.sas~data:

This data file contains the 182 person records for which PROXY, a revision of the PRXYIN data was recommended by operational specialists.

A.1.6 HCUF:

The Hundred Percent Census Unedited File provided variables:

- PFT form type
- PCMODE response collection mode
- RHHMEM respondent household member?

A.1.7 PCOMB.sas~data:

This data set combined the P-sample variables above, used for P-sample analyses.

A.1.8 ECOMB.sas~data:

This data set combined the E-sample variables above, used for E-sample analyses.

A.1.9 SIME3.sas~data:

The alternate variance computation runs accessed this dataset.

A.2 Definitions of Analysis Groups

The analyses consisted mostly of comparing subgroups of selected cases within the P sample or the E sample. Variables from A.C.E. or Census 2000 data were used to create those groups. Values in Tables 4.1.1 to 4.1.4 were generated by summing over the properly weighted and selected records in the PCOMB and ECOMB datasets.

A.2.1 Selection of Cases:

For the P-sample analyses, records of Census Day residents were selected (Inmover records had zero TESFINWT and were thus removed.), i.e. . . .

- if RPROB > 0, and
- if TESFINWT > 0.

For the E-sample analyses, only the TESFINWT > 0 criterion was needed.

For E-sample cases once thought to be surrounding block geocoding error but later found within the cluster, TESPER was reassigned:

If (TESSELECT='R' and BFUMAT='GC'), then TESPER = 0.

When only TES persons were being counted, as in numerators of statistics, they were selected in both E and P samples with: If TESPER = 1.

A.2.2 Assignment to Analysis Groups:

Many of the P-sample and E-sample groups analyzed for this report were formed directly from the levels of a variable in the estimation dataset. Table A presents the source variables for each analysis table in this report. Notes below the table elaborate the recoding definitions.

Table A. Variables Defining TES Analysis Groups

Tables	Analysis Groups	Defining Variables	Used
4.2.1-2	**Tenure**	TENURE2	as is
4.2.3-4	**Size of Metropolitan Statistical Area and Type of Enumeration Area**	MSATEA	as is
4.2.5-6	**Age and Sex**	AGESEX	as is
4.2.7-8	**Race and Hispanic Origin Domain**	DOMAIN	as is
4.2.9-10	**Return Rate Indicator**	RRATEIND	as is
4.2.11-12	**Region of the United States**	REGION	as is
4.2.13-14	**Subsampling Involvement**	WEIGHTP, WEIGHTC	recoded

Tables	Analysis Groups	Defining Variables	Used
4.2.15-16	**Type of Structure at Basic Address**	TOBA, STRCDE	recoded
4.2.17-18	**Type of Respondent**	PRXYIN, PROXY	recoded
4.2.19	**Mover Status**	MOVERPER	as is
4.2.20-21	**Imputation of Characteristics**	AMTIMP	as is
4.2.22-23	**Household Size**	POSSPSC, NESAMP	recoded
4.2.24-25	**Kinship to Reference Person**	RELAT2, POSSPSC, NESAMP	recoded

- *Defining Subsampling Involvement (Tables 4.2.13 & 4.2.14),*

If WEIGHTP not = WEIGHTC, assign to 'Subsampled';
else if WEIGHTP = WEIGHTC, assign to 'Not Subsampled'.

- *Defining Type of Structure at Basic Address in the P sample (Table 4.2.15),*

If TOBA = 1, assign to 'Single-Family Dwelling';
else if TOBA = 2, assign to 'Multi-Unit';
else if TOBA = 3 or 4, assign to 'Mobile Home';
else assign to ’Living Quarters in a Special Place and Unclassified'.

- *Defining Type of Structure at Basic Address in the E sample (Table 4.2.16),*

If STRCDE = 1, assign to 'Single-Family Dwelling';
else if STRCDE = 2, assign to 'Multi-Unit';

- *Defining Type of Respondent in the P sample (Table 4.2.17),*

One definition of the proxy group for this variable is simply PRXYIN = 1. An alternate definition, “If PRXFLG=1 or RESPNUM=99 then PROXY=’yes’” led to reassignment of 182 cases. Differences in statistics and analyses for PROXY relative to those from PRXYIN were negligible -- essentially identical at the level of rounding in these analyses.

- *Defining Type of Respondent in the E sample (Table 4.2.18),*

if (PFT in ('05','06','17','18') or PCMODE = '2') and RHHMEM in ('2','3') then PROXY=’yes’;
else PROXY=’no’;

- *Defining Household Size in the P sample (Tables 4.2.22),*

If POSSPSC = 1, assign to 'One person';
else if POSSPSC = 2 to 6, assign to '2-6 persons';
else if POSSPSC = 7 or more, assign to '7 or more persons'.

- *Defining Household Size in the E sample (Tables 4.2.23),*

If NESAMP = 1, assign to 'One person';
else if NESAMP = 2 to 6, assign to '2-6 persons';
else if NESAMP = 7 or more, assign to '7 or more persons'.

- *Defining Kinship to Reference Person (Tables 4.2.24 & 4.2.25),*

If RELAT2=9 and (POSSPSC=1 for P sample / NESAMP=1 for E sample)
assign to ‘Reference person, alone';
else if RELAT2=9, assign to 'Reference person, not alone';
else if RELAT2=1, assign to 'Spouse';
else if RELAT2 = 2 or 4, assign to 'Parent/Child ';
else assign to 'Other relatives and nonrelatives';