UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

October 30, 2008

DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-E-18

MEMORANDUM FOR Donna L. Kostanich, Co-Chair
Census Coverage Measurement Operational Integration Team

Magdalena Ramos, Co-Chair
Census Coverage Measurement Operational Integration Team

From: Richard A. Griffin *(signed)*
Chair, Census Coverage Measurement Estimation Subgroup

Prepared by: Thomas Mule
Decennial Statistical Studies Division

Subject: 2010 Census Coverage Measurement Estimation Methodology

The attached document on "2010 Census Coverage Measurement Estimation Methodology" is being provided as background material for the Census Coverage Measurement Estimation Workshop to be held in January 2009.

Attachment

cc:
DSSD CCM Contacts List

2010 Census Coverage Measurement Estimation Methodology

			Page
	Introduction		1
1.0	Universes		2
2.0	Coverage Estimation for People in Housing Units		2
	2.1	Estimation of Net Error for People in Housing Units	2
		2.1.1 Dual System Estimation Model	2
		2.1.2 Determining Correct Enumerations	3
		2.1.3 Determine the Proportion of People Correctly Enumerated	6
		2.1.4 Treatment of Movers	7
		2.1.5 Dual System Estimation Using Logistic Regression	7
		2.1.6 Correlation Bias Adjustment for People in Housing Units	12
		2.1.7 Population Estimates Adjusted for Correlation Bias	14
	2.2	Component Estimation for Persons in Housing Units	14
		2.2.1 Four Coverage Components for Persons in Housing Units	15
		2.2.2 Types of Erroneous Enumerations	17
		2.2.3 Enumerated in Wrong Location	19
		2.2.4 Estimation of Three Components for Census Records	19
		2.2.5 Estimation of Omissions	22
	2.3	Estimation Procedures for Coverage of Persons in Housing Units	24
		2.3.1 Characteristic Imputation (E and P Sample)	25
		2.3.2 Noninterview Adjustment	25
		2.3.3 Unresolved Correct Enumeration Status for Net Error (E sample)	26
		2.3.4 Unresolved P-sample Inclusion Status for Net Error (P sample)	29
		2.3.5 Unresolved Match Status for Net Error (P sample)	32
		2.3.6 Unresolved Enumeration Statuses for Component Error (E sample)	34
		2.3.7 Final Probabilities for Net Error	36
		2.3.8 Final Probabilities for Component Error (E sample)	37
		2.3.9 Weight Trimming	40
		2.3.10 Logistic Regression Models for Net Error	40
		2.3.11 Correlation Bias Adjustment for Persons in Housing Units	42
		2.3.12 Final Estimates for Net Error	42
		2.3.13 First-Stage Ratio adjustment for Component Error	43
		2.3.14 Final Estimates for Component Error Including Second-Stage Ratio Adjustment	43

			Page
3.0	Coverage Estimation for Housing Units		43
	3.1	Estimation of Net Error for Housing Units	43
		3.1.1 Dual System Estimation Using Logistic Regression	44
	3.2	Component Estimation for Housing Units	45
		3.2.1 Types of Erroneous Enumerations of Housing Units	47
		3.2.2 Estimation of Components for Census Housing Units	48
		3.2.3 Estimation of Housing Unit Omissions	50
	3.3	Estimation Procedures for the Coverage of Housing Units	50
		3.3.1 Characteristic Imputation (E and P Sample)	50
		3.3.2 Unresolved Correct Enumeration Status for Net Error (E Sample)	51
		3.3.3 Unresolved Housing Unit Status for Net Error (P sample)	51
		3.3.4 Unresolved Match Status for Net Error (P sample)	52
		3.3.5 Unresolved Enumeration Statuses for Component Error (E sample)	52
		3.3.6 Final Probabilities for Net Error	53
		3.3.7 Final Probabilities for Component Error	54
		3.3.8 Weight Trimming	55
		3.3.9 Logistic Regression Models for Net Error	55
		3.3.10 Final Estimates for Net Error	57
		3.3.11 First-Stage Ratio Adjustment	57
		3.3.12 Final Estimates for Component Error including Second-Stage Ratio Adjustment	57

Introduction

The Census Coverage Measurement (CCM) program will evaluate the coverage of the 2010 census and has been charged with providing information that can be used to improve future censuses. This document continues the discussion started in the 2010 Coverage Measurement Executive Brief (Singh, 2003). The brief documented and communicated a) what we hoped to achieve in 2010, b) the Census Bureau's goals at that time and c) what we would have ideally liked to accomplish. In November 2005, the Census Bureau released an updated plan after addressing the implications, methods, and resources necessary to accomplish these goals (Singh, 2005).

The major goals of the CCM program are:

- begin producing measures of coverage error, including its components of erroneous enumerations and omissions,
- produce these measures of coverage error not only for demographic groups and geographic areas, but also for key census operations
- continue to provide measures of net coverage error.

The previous documents addressed <u>what</u> we would like to achieve and <u>not</u> as much how we are going to do it. The documents also stated that the Census Bureau recognizes that we may not be able to fully achieve all of these goals and objectives laid out. There may be limitations in terms of resources or amount of sample that will be allocated to this effort. Also, there may be other inherent operational or technical limitations. The CCM Design document describes in some degree of detail the design of other aspects of coverage measurement, such as sampling, data collection and matching (Moldoff 2008).

This document continues the discussion by starting to focus on <u>how</u> we are going to achieve these goals from an estimation perspective. Wherever possible, we will show how our current approach for 2010 compares or differs from the original Accuracy and Coverage Evaluation (A.C.E.) or the A.C.E. Revision II estimation methodologies. We will also identify any previous goals in the 2003 Executive Brief that are no longer being considered. Some of these changes were already identified in the updated plan (Singh, 2005).

This document focuses on the coverage of the census population in the United States. It provides more detailed information on the estimation than noted in the high-level CCM overview (Whitford 2008a) and the CCM Design document. This estimation methodology is designed to support the coverage statistics and estimation domains in Whitford (2008b). While not specifically addressed here, the coverage measurement of Puerto Rico will follow a similar methodology.

Section 1 lists what universes will be evaluated by the CCM program for the 2010 census. Section 2 documents the estimation methodology for the coverage of people in housing units. Section 3 documents the estimation methodology for the coverage of housing units.

This document focuses on the point estimation. The CCM Estimation team is researching how to use replication techniques to measure the standard errors of the point estimates due to sampling variance and also account for missing data variance as well.

1. Universes for Coverage Measurement

The CCM program will be evaluating the coverage of a) persons in housing units and b) housing units. We are not evaluating the population in group quarters. We will **no longer** be evaluating coverage of group quarter facilities which was one of the goals listed in the Executive Brief because of our resource limitations. The November 2005 updated plan documented this change. Similar to the 2000 A.C.E., any area designated as Remote Alaska enumeration area will be out-of-scope as well for the coverage evaluation. Also similar to 2000, the population in transient living situations like those experiencing homeless will not be evaluated by the CCM program. Thus, we will be evaluating the same universes as the 2000 A.C.E.

2. Coverage Estimation for People in Housing Units

This section describes the coverage estimation for the universe of people in housing units. Section 2.1 provides an overview of the estimation of net error. Section 2.2 gives an overview of the estimation of component error for the 2010 CCM. Finally, Section 2.3 provides details on the estimation procedures and methods that will be used for both the net error and component error estimation.

2.1 Estimation of Net Error for People in Housing Units

We will estimate net error by comparing the estimate of the true population from the CCM program to the census count. This will be done by the following:

$$Net\,Error = True\,Population - Census \quad (1)$$

2.1.1. Dual System Estimation Model

We will be implementing dual system estimation to estimate the true population of people in housing units. There is a long history of using dual system estimation in measuring coverage errors in a census (Marks, 1979; Wolter, 1986; U.S. Bureau of the Census, 1985; Hogan 1992, 1993; Hogan 2003; U.S. Census Bureau 2004).

The standard Petersen (1896) or Sekar-Deming estimator assuming independence is expressed as:

$$N_{++} = N_{+1}\left(N_{1+}/N_{11}\right) \quad (2)$$

Where N_{++} is the total number of people,
N_{+1} is the number of people counted in the census
N_{1+} is the number of people counted in the survey
N_{11} is the number of people counted in both the census and the survey

2.1.2 Determining Correct Enumerations

This first step in operationalizing the dual system estimator is to define and estimate the list or set of individuals "correctly" in the census. The determination of the correct enumeration status is done since the general net coverage model assumes that spurious events like duplications, nonexistent cases or out-of-scope cases, such as an individual born after Census Day, have been identified in both systems and accounted for in the estimation (Wolter 1986). In this context "correctly" has four dimensions:

1. Appropriateness
2. Uniqueness
3. Completeness
4. Geographic correctness

"Appropriateness" means that the person should be included in the census. People who die before or who were born after the census reference date (April 1 in the U.S.) are not part of the population (universe) to be measured. Similarly, records that refer to fictitious "people," tourists or animals are out-of-scope.

"Uniqueness" refers to the fact that we wish to measure the number of people included in the census, not the number of census records. If more than one record refers to a single person, the count of records must be reduced for purposes of the dual system estimation.

"Completeness" means that the census record must be sufficient to identify a single person. If it lacks sufficient identifying information, we cannot determine for all of those records whether the person was appropriately and uniquely included in the census, nor can we determine whether he or she was also included in the survey.

Although completeness is necessary for the dual system estimation, the census count includes imputations and other incomplete enumerations. Census operations normally have a requirement for a "data-defined person." In the 2010 Census, the requirement is two characteristics where the name counts as a characteristic. Name must have at least three characters in the first and last name together. The characteristics that are included in the counting are relationship, sex, race,

Hispanic origin, and either age or year of birth. When a record does not meet these requirements census processing substitutes (imputes) a data-defined record. Since the census processing flags all these whole-person imputations, these records can be identified in the census files.

Additionally, there are person records that are acceptable for census processing but insufficient for use in the dual system estimation. This group includes records with reasonably complete data but without a person's name. Accurate matching or additional interviewing is not possible for all of these cases[1]. For the dual system estimation, the definition for "sufficient information for DSE processing" is complete name and two characteristics.

"Geographic correctness" means that people are included in the census where they should be included. Enumerations outside the defined search area that are counted in the census are not considered correctly included in the census for dual system estimation. This area must be searched during the matching process as well as searched for duplication. As the size of the small area increases, the complexity increases and the chance of false matches grows. Normally a small area is defined such as a block cluster and the surrounding ring of collection blocks. A block cluster is a single collection block or a group of contiguous collection blocks. Two dimensions must be defined to operationalize a smaller area: (1) correct location and (2) the search area around the correct location. The "correct location" defines where, under the Census residence rules, the person should be included.

The second dimension of geographic correctness is the area of search around the correct location, i.e., the search area. The concept of a search area is to accommodate errors in either the census or survey assignment of residents to a particular geography. It has the effect of lowering the variance and can, in some circumstances, lower the bias as well. The CCM uses the following definition:

> A person was correctly enumerated if the person was counted in the block cluster containing his/her Census Day residence; or
>
> if he/she was included by the census in a housing unit that is his/her Census Day residence and the housing unit was included in a block adjacent to the correct block cluster.

The definition of "correctly included" does not depend on the person's characteristics being correct but only on the location of his census residence. For example, if a person was really 19 years old but was counted in the census as 17, the person is still considered as correctly included.

---

[1] During clerical matching to support the component estimation, we are able to determine for some of these insufficient cases whether they were correct or erroneous based on results of matching people collected during the CCM Person Interview to these cases. These clerical matching results are not used for dual system estimation so that we do not introduce matching error.

To estimate the number of people correctly included in the census, one must take a sample of all data-defined census enumerations. A sample of census person and housing unit records, referred to as the Enumeration (E) sample, are taken to determine how many of them were correctly counted in the right search area. This sample is called the enumeration (or E) sample. Census whole-person imputations are not part of the E-sample person frame.

To maximize correlation with the Population (or P) sample, the CCM first defines a set of sample areas. The sample areas are the block clusters. The P sample is the second independent sample for the dual system estimation. If a block is sampled, all census records coded to that block, even incorrectly, fall into sample. If the block contains many census housing unit records, it may be subsampled.

The records in the E sample are processed for completeness. Only records that meet the minimum completeness requirement can be considered as correctly enumerated for dual system estimation. Records are searched throughout the country to see if the person was counted more than once. The duplicate search is done by using computer matching and computer-assisted clerical matching. If more than one record is found, the CCM determines which is correct and which are erroneous.

Appropriateness and geographic location cannot be determined from the census enumeration alone, but require additional interviewing. If interviewing locates a member of the household, or an acceptable respondent who can confirm the person's existence and that the person had his/her Census Day residence there on April 1, the enumeration is accepted as correct.

The followup of the E sample in the CCM is necessary because experience has shown that the census contains an appreciable level of erroneous enumerations. This was shown in the A.C.E. for Census 2000. Ideally, this source of error would be controlled and excluded in the enumeration phase but this approach has not proven practical. Some countries assume that census erroneous enumerations can be ignored.

If the respondent reports that the person did not live in the block or search area on April 1, the enumeration is classified as erroneous. This can occur when the person responded to the census but moved before April 1; the person moved in after April 1 but was enumerated at another address; or when a parent incorrectly reports a college student as living at home.

The interviewers may determine that the person never existed or was never associated with the block. These records are considered erroneous for dual system estimation. It can be difficult in some cases to prove that a "person" was not real, especially in a large block.

An important source of error arises from the need to often accept proxy responses to verify many enumerations. If the proxy reports a different "correct" residence than the person himself would, an enumeration could be miscoded. The CCM uses proxy interviews for households that have moved between the time of the census and the time of the CCM interview. The CCM requires

the interviewers to find at least three knowledgeable respondents before coding a record as fictitious. However, since the person might have lived somewhere else in the block, it can be difficult in some situations to code the fictitious people correctly.

2.1.3. Determining the Proportion of People Captured in the Census

Having defined the set of correctly enumerated people, the next step in dual system estimation is to estimate the census coverage rate N_{11}/N_{1+}. This rate is also called the match rate.

The CCM program will independently list the housing units in the sample of block clusters. A sample of these listed housing units will be selected and those will be the P-sample housing units. We will then conduct the CCM Person interview at those housing units. The interviewing will result in a list of people who should have been enumerated in a housing unit in the census. This list constitutes the population or P sample.

The CCM will then attempt to match the P-sample people to the census records. A case can only be matched if the census record is located in the correct search area. The correct search area is the block cluster and the surrounding ring of collection blocks where the person or housing unit should have been counted in the census according to the census residence rules. Dual system estimation assumes that the P-sample people and housing units can be matched correctly to the census results and this determination is made with no error (Wolter 1986).

Conceptually, estimating the match rate entails (1) taking a sample of people, (2) determining whether they should be enumerated in the census, and (3) determining whether they were, indeed, correctly enumerated, using the same definitions as were used to determine the correct enumerations in the previous section. If an unbiased sample can be drawn of people who should have been enumerated and if we can determine whether they actually were correctly enumerated, then the dual system estimation will produce asymptotically unbiased estimates. If each step can be approximately correct, the results will approach an unbiased estimate.

The CCM guards against unnecessarily introducing operational dependence by applying the same requirements on P-sample records that were applied to determine the correct enumeration status of the E sample in the previous section. The P-sample records are processed on the same four dimensions listed in Section 2.1.2. Out of scope records are screened out. Occasionally, survey duplicates occur and these are removed. Finally if the survey interview does not meet minimal standards the case is converted to a nonresponse and is handled by missing data methods.

The matching and followup result in the ability to estimate the number of people correctly counted in the survey (N_{1+}) and the number of people correctly counted in both the census and the survey (N_{11}). It is not the two estimates that are of specific interest but the ratio of the two for the dual system estimation.

2.1.4. Treatment of Movers

People who move between the census reference date and the time of the survey interview present a challenge for designing a dual system estimation for a census application. First, people who move are more likely to be missed by the census and by the survey. Secondly, if a person has a different residence at the time of the survey than he did at the time of the census, one must decide where to sample him.

The 2010 CCM is doing a variation of how movers were treated for the 1990 PES. In the 1990 PES, movers were sampled where they lived at the time of the survey interview (nonmovers and inmovers since Census Day). We then searched the census records at their April 1 residence according to Census residence rules. This is known as procedure B (Marks 1979). For the 2010 CCM, we are also including people who moved out of the housing unit since Census Day but who have no probability of selection for Interview Day. One example of this is a high school senior who is enumerated at home on April 1st and then resides in a college dorm in September. These people are being included in the P sample since they were living in a housing unit on Census Day. Since they have moved to a group quarters, they have no chance of being selected in sample since group quarters are out of scope for the CCM coverage evaluation. This treatment is being called "Procedure or PES-B+."

Matching the inmovers to the enumerations at their Census Day residence search area has some implications. The inmovers can only match to correct enumerations. First, the enumeration must be in the correct geographic search area =-based on the Census Day address provided. The same search area around the correct location listed in Section 2.1.3 is applied to this address. Second, the enumeration must meet the completeness requirement. During the A.C.E. matching, the clerks were able to use both a) the name and characteristic data captured on the Census Unedited File (CUF) and b) the scanned census forms to determine if the case met the completeness requirement. Since A.C.E. used PES-C, nonmovers and the outmovers were only matched within the sample block cluster search area. For the CCM, the scanned census forms will only be available for the sample block cluster areas and thus not necessarily available for the inmover address search areas. To avoid a bias, a consistent application of completeness is necessary. Therefore completeness of census records will be based only on the data captured on the CUF for the dual system estimation.

2.1.5 Dual System Estimation Using Logistic Regression

The dual system estimation requires the assumption that census capture probabilities be independent for all individuals on a list (Wolter 1986). The independence assumption can fail either due to causal dependence between the Census enumeration and the coverage survey enumeration or due to heterogeneity in capture probabilities. Causal dependence occurs when the event of an individual's inclusion or exclusion from one system affects their probability of inclusion in the other system. However, even if causal independence is true for all individuals, the independence assumption can be violated by heterogeneity of inclusion probabilities across

different individuals. If the census inclusion probabilities are not equal for all individuals, the heterogeneity results in correlation bias. Griffin (2000) provides details on the impact when this assumption fails.

The previous dual system estimates in 1980, 1990 and 2000 used post-stratification to minimize the impact of the failure of these assumptions (Chandrasekar and Deming 1949) . Post-stratification or grouping of individuals likely to have similar census inclusion probabilities and calculating dual system estimates within post-strata is done to decrease heterogeneity and thus decrease correlation bias. Post-stratification results in estimating the population of each stratum and aggregating the results over all of the strata The post-stratification approach has some significant limitations. It limits the number of factors that can be included because each factor added to the post-stratification can crudely be thought of as cutting the post-stratum sample sizes in half. This is due to the implicit estimation of many high-order interactions across factors in the post-stratification. For more information on the post-stratification approach to A.C.E. estimation, see Appendix A and U.S. Census Bureau (2004).

Haberman, Jiang and Spencer (1997) researched using logistic regression models instead of post-stratification to generate dual system estimates of the 1990 PES. Their research showed that separate models could be fit to determine the correct enumeration status of the E sample and the match status of the P sample. The results of these models could be used to generate population estimates. Griffin (2005) introduced the plans for several alternative net error estimation alternatives using these logistic regression ideas. Each plan has advantages and disadvantages to the total population estimates and estimates for specified sub-populations.

For the 2010 CCM, we will use the estimator referred to as "N_0" in Griffin (2005). This estimator has the form:

$$N_O = \sum_{j \in C} \pi_{dd,j} \frac{\pi_{ce,j}}{\pi_{m,j}} \quad (3)$$

Where $\pi_{dd,j}$ is the predicted probability census case j is data-defined,
$\pi_{ce,j}$ is the predicted probability that the census case is a correct enumeration
$\pi_{m,j}$ is the predicted probability that the census case was matched
j is the a census enumeration and
C is the estimation domain (total, tenure status, age/sex groupings, etc.).

We have chosen this estimator for several reasons:

- Some of the estimators (N_2, N_3, N_{2R} and N_{3R}) used weighted summations of the E sample instead of using a summation over all of the census cases (N_0 and N_1). These estimators could have allowed information that was collected only in the sample areas to be used in the estimation. However, the estimators that use only sample data impose limitations on

what small area estimates could be generated and would have required a synthetic estimation plan to be developed for some of the estimates.

- Two estimators (N_0 and N_1) involve summing the predictions over all of the census records. The difference was how the data-defined status was handled. The census processing is able to determine this status for every census record. The N_0 estimator uses a predicted rate of being data-defined based on a model while the N_1 estimator replaces the predicted rate by an indicator value of the data-defined status of the census record in the estimation. Why would we use an estimator that uses a model probability prediction when the resolved status is available for every case? The reason is that the N_0 and N_1 estimators produce negligible differences for the population estimates for large estimation domains. However, the N_1 estimator has a potential weakness for any small estimation domains that have relatively small number of data-defined cases. The N_1 estimator may produce estimates that are too low unless the match rate model contains a covariate that can balance for the small number of data-defined cases. Mulry et al. (2008) showed this result when using the N1 estimator to generate block cluster-level estimates for clusters with low data-defined rates that did not include a covariate to balance the match rate model. The match status model would need to account for all pre-specified small area domains and also be robust to produce estimates for small area domains requested later. Since the N_0 estimator uses a model for all three of the predictions, it has robust properties to provide some protection against possible balancing errors.

The 2010 estimation will use logistic regression instead of post-stratification. This approach was considered based on the thought that this approach would offer more flexibility and possibilities for reducing sampling error, correlation bias and synthetic error in the estimates. Our research has shown potential reductions of synthetic error by allowing the inclusion of more variables as main effects. The models allow us to utilize more variables in the model and gives us a better way to generate estimates of percent undercount for individual years of age. Our research in using these models to reduce sampling error or correlation bias has not been as beneficial as initially thought.

Logistic regression uses a statistical modeling approach to be able to estimate the relationship of a binary outcome to independent variables like race, age, sex and others. In logistic regression, probabilities of outcome categories are assumed to be a function of a linear combination of the explanatory variables. This function is also called a link function. Logistic regression uses the logit link function. The independent variables may be continuous, discrete, or a mix of any of these as compared to post-stratification cells where only groupings or categorical variables can be used.

One requirement of the production logistic regression processing is that a standard statistical package like SAS® be utilized when running the logistic regressions. These packages allow weights to be utilized when solving the weighted maximum likelihood estimates of the regression coefficients. These packages already have written and tested the code to implement

the necessary algorithms to perform these regressions. These packages also generate and report a set of diagnostic output about the model fitting. These packages also provide information if errors are encountered like an independent variable or variables are not being coded correctly.

For dual system estimation of people in housing units, we need to run three logistic regression models to relate each of the outcomes to the chosen independent variables. When sample data is used, the appropriate sampling weights and probabilities need to be used in the logistic regression. Section 2.3 provides details on the estimation procedures for missing data, weight trimming and dual system estimation. Section 2.3 also provides details on estimation procedures on how the logistic regression models account for the final sampling weights and assigned status probabilities.

The Estimation Team is currently researching which independent variables to use in the model. When running each model, we are estimating a coefficient parameter for each variable. The parameter measures how each variable impacts the probability of an outcome happening in conjunction with the other variables being in the model as well. See Appendix B for more information on logistic regression.

The three outcomes to be modeled by logistic regression are:

1. Was a census case data-defined?

2. Was the data-defined E-sample case a correct enumeration in the correct location for net error estimation?

3. Did a P-sample case match to a census enumeration in the correct search area?

For each of the three outcomes, Table 1 summarizes the successful outcome of the dependent variable , data utilized and whether weights will be utilized when running these three logistic regression models. The data-defined regression will use all of the census persons in housing unit records in the regression while the other two will use sample data. This is similar to the post-stratification estimates for the 1990 PES and March 2001 A.C.E. estimates where the data-defined rate for each post-stratum was determined using the entire census data.

For the other two regressions, we will use the sampling weights in the correct enumeration and match status regressions for those sample data. When running these models, the independent variables will be main effects or interactions of those effects. Our current plan is to include the same main effects in each model but not necessarily have the same interactions in each model.

Table 1: Logistic Regression Model Information

Successful Outcome of Dependent Variable	Data Source	Weights Used in Regression
Data-Defined	All Census People in Housing Units	No
Correct Enumeration	E-sample People	Yes
Match	P-sample People	Yes

We will use the results of the three logistic regression models to make predictions of the probability of being a data-defined enumeration, probability of being a correct enumeration and probability of being matched shown in equation 3. We can make the predictions by using a) the values of the independent variables of the census case and b) the estimated parameters of the respective models shown above. The independent variables will be the same as those used in the logistic regression models that generated the estimated parameters.

The general predicted rate formula is:

$$\pi_j = \frac{1}{1 + e^{\left(-\beta_0 - -\sum X_{ij}\beta_i\right)}} \qquad (4)$$

Where π_j is the predicted rate for census person j,
X_{ij} is the value of the *i*th independent variable[2] in the census person j and
β_i is the parameter estimate of the intercept or the ith independent variable from the respective logistic regression.

Based on the three prediction results for each census case *j*, we can then compute a preliminary dual system estimate for each census case j. This is preliminary since it will be adjusted for correlation bias in a future step.

$$PREDSE_j = \pi_{dd,j} \times \frac{\pi_{ce,j}}{\pi_{m,j}} \qquad (5)$$

Where PREDSEj is the dual system estimate prior to correlation bias adjustment for census person j,
$\pi_{dd,j}$ is the predicted data-defined rate for census person j,
$\pi_{ce,j}$ is the predicted correct enumeration rate for census person j and
$\pi_{m,j}$ is the predicted match rate for census person j

[2] The general formula shows a summation over *i* independent variables. Our current plan is to have the same main effects in each model but the interactions could be different if warranted.

2.1.6 Correlation Bias Adjustment for Persons in Housing Units

Correlation bias exists whenever the probability that an individual is included in the census is not independent of the probability that the individual is included in the CCM or there is heterogeneity in the capture probabilities. This form of bias generally has a downward effect on estimates, because people missed in the census may be more likely to also be missed in the CCM.

Estimates of correlation bias will be calculated using the "two-group model" and sex ratios utilizing Demographic Analysis data. Demographic Analysis is another evaluation of the net census coverage done by the Population Division. Demographic Analysis is a macro-level approach utilizing various demographic data such as administrative statistics on births, death, legal international migration and Medicare enrollments as well as estimates of legal emigration and unauthorized imigration. Robinson and Adlakha (2002) document the Demographic Analysis estimates for 2000. Judson (2008) documents the tentative plans and research ideas for the 2010 Demographic Analysis program.

The sex ratio is defined as the number of males divided by the number of females. The two-group model assumes no correlation bias for females over 18. It also assumes that Black males over 18 have a relative correlation bias that is different than the relative correlation bias for non-Black males.

The choice of doing separate adjustments for Black and non-Black populations is based on the previous availability of Demographic Analysis information for those two groupings. The Black population will be identified by any census enumeration who reported their race as Black or African American (alone or in combination with other races). The non-Black population will be any census enumeration that did not report a race of Black or African American.

The Demographic Analysis program will provide the CCM estimation with their population estimates of the Black and non-Black populations by gender for each individual year of age. Since the Demographic Analysis estimates are for the entire population in the United States, we will need to subtract out the census counts of the Black and non-Black populations enumerated in group quarters. The CCM program will also need to subtract out the population in the type of enumeration area of Remote Alaska. Remote Alaska is out-of-scope for the 2010 CCM. This will provide estimates of the population in housing units from which sex ratios can be calculated.

We are planning on doing the adjustment by individual years of age for the 18+ Black and 18+ non-Black male populations. This is different than the adjustments in A.C.E. Revision II that used three age categories: 18-29, 30-49, and over 50. Shores (2003) documents the correlation bias adjustment for the A.C.E. Revision II estimates. Bell (1999) documents using the two group method and age groupings to adjust logistic regression dual system estimates of the 1990 PES data as part of the 2000 A.C.E. research.

The correlation bias adjustment is done so that the final dual system estimation sex ratios will agree with the Demographic Analysis-adjusted sex ratios for Blacks and non-Blacks. The correlation bias adjustment is calculated for the race/age groups by the following. This factor will be used to adjust the male population in each of the race/age groups.

$$c_k = \frac{\sum_{j \in Female \cap k} PREDSE_j}{\sum_{j \in Male \cap k} PREDSE_j} r_{DAk} \quad (6)$$

Where c_k is the correlation bias adjustment for males in the kth race/age group,
$PREDSE_j$ is the preliminary DSE prior to correlation bias adjustment and
r_{DAk} is the Demographic Analysis-adjusted sex ratio from adjusted of the kth race/age group.

The following are possible refinements being considering that may be made to the methodology to improve this adjustment:

- implement an adjustment for the under 18 population. The A.C.E. Revision II estimated an overcount for the 0 to 9 population while Demographic Analysis estimated an undercount.

- implement an adjustment for Hispanic population if Demographic Analysis data is available. There is a possibility that the Demographic Analysis program may have estimates for Hispanics the population age 30 and under in 2010.

- smoothing of the Demographic Analysis sex ratios. This smoothing can be done either before or after the census counts for out-of-scope populations (group quarters and Remote Alaska) is performed.

2.1.7. Population Estimates Adjusted for Correlation Bias

We will adjust our dual system estimator shown in equation 3 so that we can minimize the correlation bias for the male population. This adjustment is based on results of the previous section.

The estimator of the total population with the correlation bias adjustment is:

$$DSE = \sum_{j \in C} PREDSE_j \times CB_j \quad (7)$$

Where CB_j is the correlation bias adjustment factor for males in the kth race/age group or 1 for all other cases.

If we decide to implement a correlation bias adjustment for any of the population under 18 then those census cases will receive a correlation bias adjustment factor based on the determined methodology.

2.2. Component Estimation for Persons in Housing Units

For 2010, the CCM program has been given the objective of measuring the components of census coverage error in addition to net error. Since the primary focus of the previous coverage evaluations of the decennial censuses has been on net error estimation, the Census Bureau only released estimates of erroneous enumerations and omissions at a national level (Barron 2001). These estimates were based on available data from the matching operations and several simplifing assumptions. These estimates were not broken down by race/ethnicity, age/sex or other groupings. The estimates were only point estimates and did not have any accompanying standard errors reported. After the errors were found and the A.C.E. Revision II estimates were completed, no new estimates of components were released because of the limitations in the data.

The CCM program has developed an estimation methodology to produce estimates of component error that will evaluate the 2010 Census and provide information to help with the planning for the 2020 Census. This plan allows for different types of estimation domains with the initial proposals in Whitford (2008b).

The CCM program has also expanded the items of component error to include more than just erroneous enumerations and omissions. Section 2.2.1 documents the four items that will be reported as components of census coverage error. We believe that by analyzing these four items together we will provide a more thorough evaluation of the 2010 Census.

The remaining parts of Section 2.2 further document the component estimation. They provide a high-level overview of the data and estimation for components. Section 2.3 provides more details on the estimation processing for components.

For dual system estimation just shown in Section 2.1, we have data available based on dual system application that could allow us to estimate quantities of erroneous enumeration using the E-sample data and nonmatches using the P-sample data. The naive assumption is that these estimated quantities can directly be used to generate estimates of erroneous enumerations and omissions for component estimation. However, these estimates of erroneous enumerations and nonmatches are based on the stringent rules of how we implement dual system estimation that, if used, overstate the estimate of erroneous enumerations and omissions at a national level. One example of the difference is that an enumeration may be erroneous for dual system estimation because it was enumerated in the wrong location but if the person was enumerated once and only once then that enumeration would be correct for national estimates of components. Using the estimates of erroneous enumerations from dual system estimation will overestimate the number of erroneous enumerations nationally. Because of this, the CCM has expanded the interviewing, matching and followup operations to provide the necessary information for both dual system estimation and component estimation.

2.2.1 Coverage Error Components for Persons in Housing Units

For 2010, the CCM is defining four items as components of census coverage error for persons in housing units. The first three of the four are the possible classifications of the census enumeration of people in housing units. The CCM program wanted to expand beyond the reporting of erroneous enumerations in the census as was done in the past. The fourth component is omissions of people that were not included in housing unit enumerations in the census.

The four components of census coverage error for persons in housing units are:

1. Correct Enumerations

   The CCM will estimate the number of correct enumerations in the final census count. At the national level, an enumeration will be considered correct if it was counted once and only once in the housing unit universe in the United States. At a national level and for evaluating estimation domains like race/ethnicity domains, if the person was supposed to be enumerated in a housing unit and was included in a housing unit anywhere in the United States then they will be considered correct. If such a person or unit was included multiple times, one of the enumerations will be considered correct and the other enumerations will be erroneous. For net error estimation in order to be considered correct, the enumeration needs to meet the requirements shown in Section 2.1.2. Component estimation does not require that the data-defined cases in the E-sample cases have sufficient amount of data collected for dual system estimation.

2. Erroneous Enumerations

The CCM will estimate the number of erroneous inclusions in the final census count. Erroneous inclusions are duplicate numerations or enumerations that should not have been counted in a housing unit.

In a separate tabulation, we are also estimating enumerations being included in the wrong location. While we will be treating these as correct in the other tabulations, we will be producing results that show the estimates of wrong location by varying geographic distances. The person matching and interviewing can attempt to determine if the person should have been enumerated in a housing unit somewhere else outside of the sample block cluster search area. The CCM will identify whether the person should have been counted in the 1) same county but outside of the block cluster search area, 2) different county in the same state or 3) different state. The results will be summarized to estimate the number of resulting cases for these three geographic definitions to the national level. Results will not be provided for individual states or counties. Section 2.2.3 provides some more information on this estimate.

3. Whole-Person Census Imputations

The CCM program will tally and report the number of whole-person census imputations in housing units. The CCM program will not be evaluating if imputations were correct or erroneous.

Table 2 shows the five groupings of the types of Whole-Person Census Imputation:

Table 2: Whole-Person Census Imputations by Category

Count Imputation 1. Status Imputation - No information about the housing unit; Imputed as occupied, vacant or non-existent. Those imputed as non-existent are removed from the census files. 2. Occupancy Imputation - Existence of housing unit confirmed but no information as to occupancy status; Imputed as occupied or vacant. 3. Household Size Imputation - Occupied status confirmed but no information as to household count; the household count is imputed.
Whole-person Characteristic Imputation 4. Substitution - Population count known; all characteristics imputed for the entire household 5. Totally Allocated - Population count known; all characteristics imputed for some, but not all, persons in the household.

Note: Any housing unit imputed as occupied during count imputation will also have its household count imputed, which results in whole-person imputations.

4. Omissions

The CCM program will estimate the number of omissions of people in housing units from the census. The CCM program is also being asked to determine how many of the person omissions were because the housing unit was not included in the census.

2.2.2 Types of Erroneous Enumerations

For people in housing units, there are several reasons why a case can be an erroneous enumeration. In addition to overall estimates of erroneous enumerations, the CCM program wants to provide estimates by the reason why the case was considered erroneous. We believe that knowing the reason for the erroneous inclusions will provide more information about how to improve the 2020 Census.

Table 3 documents the two groupings of erroneous enumerations that we plan to estimate. The first group are people who should not have been enumerated in the housing unit universe at all. These reasons include a) discrepant or fictitous b) should have been enumerated in a group quarters but was not, c) born after Census Day, d) died before Census Day, and e) should not have been counted in the United States according to the Census residence rules.

The reasons for the combination into one group are because:

- Some of the reasons (e.g., should have been enumerated in group quarters, born after Census Day, died before Census Day) happen so infrequently that there is not a unique match code for these reasons. To minimize matching error, we do not want to implement clerical procedures for infrequently occurring events. These are identified by an overall match code about being erroneously included since they are all reasons for being out-of-scope for the CCM.

- The CCM examined the keyed data from the Person Interview and Person Followup from the 2006 CCM Test. The data were analyzed to see if these infrequently occurring reasons could be identified. If the questions were answered then some of the cases could be identified. However, if these erroneous enumerations were identified based on notes provided by the interviewer then these cases were not able to be identified.

- The CCM would need to develop a missing data mechanism for the unresolved cases that supports these infrequently occurring outcomes.

- The Census Bureau has minimum sample size requirements for reporting results so some of these may have had to have been collapsed together because of small sample sizes.

Table 3: Types of Erroneous Enumerations for People in Housing Units

Overall Type	Subtype	Description
Persons Who Should Not Have Been Counted in Housing Unit Universe At All	Discrepant	This record is fictitious and does not correspond to an actual person.
	Resides in GQ	The person was enumerated in the housing unit universe. The person should have been *but was not* enumerated in the group quarters universe.
	Born After Census Day	The person was born after Census Day.
	Died Before Census Day	The person died before Census Day
	Other Reasons	The person is a U.S. citizen and on Census Day is working, studying, or living abroad or the person is a foreign visitor who is temporarily visiting the United States.
Duplication	To other enumerations in housing unit	The enumeration is a duplicate enumeration of a person who was counted correctly in a housing unit in the census. Estimates will be generated by groupings of the geographic distance of the duplication: 1) same sample block cluster search area, 2) same county and state, 3) different county in the same state and 4) different state[1]
	To an enumeration in a group quarters	Some people will be enumerated correctly in a group quarters and also be erroneously included in a housing unit. The erroneous housing unit enumeration will be classified as a duplicate. Estimates will be generated by groupings of the type of group quarters. Examples include college dormitories, nursing homes, correctional institutions, juvenille institutions and others

[1] any duplicates in the last three groupings are also outside the sample block cluster search area.

The second group are the duplicate enumerations. The Further Study of Person Duplication (FSPD) estimated at least 5.8 million duplicate enumerations in the 2000 Census (Mule 2003). That study showed estimates of duplication to people in other housing units by different geographic distances (in block cluster search area; in same county; in different county same state; and in different state). The FSPD showed estimates of duplication to those people also enumerated in a group quarters by type of facility. As stated in Whitford (2008b) and in Table 3 of this document, the CCM plans to produce similar estimates for the 2010 CCM.

2.2.3. Enumerated in the Wrong Location

Some people will be correctly included in the housing unit universe but will be enumerated in the wrong location. For the estimation and reporting of the four types of components listed in Section 2.2.1, these enumerations will be classified as correct enumeration. However, the CCM wanted to assess the magnitude of these occurrences. These estimates focus on a person indicating that they should have been enumerated somewhere else. These are not an estimate of people in housing units and the housing unit was assigned to the wrong collection block (geocoding error).

For our evaluation of the census, we needed to determine how far away the enumeration should have been enumerated to be considered in the wrong location. If a person should have been enumerated in the next collection block then there will probably be a very small amount of error for census tabulations. For our evaluation, we decided that the wrong location was any housing unit that was located outside of the search area for dual system estimation. The search area is the collection blocks of the sampled block cluster and the one ring of collection blocks around it.

For this estimation, we will generate estimates by three groupings of geographic distance:

- wrong location but still in the right county and state
- wrong location is in a different county in the same state
- wrong location is in a different state

2.2.4 Estimation of the Three Components for Census Person Records

Section 2.2.1. documented the following three components of census coverage based on our evaluation of the enumeration of people in housing units in the census:

- Correct Enumerations
- Erroneous Enumerations
- Whole-Person Census Imputation

This section provides the estimation methodology to generate the estimates for an estimation domain. Whitford (2008b) documents our current plans for the estimation domains that we are planning to evaluate for component estimation. Examples of estimation domains are the overall estimate, the American Indians on Reservation domain, Renters and 18 to 29 year-old males. We are preparing to produce rate and level estimates. We have also developed our methodology to be able to produce estimates if other estimation domains are requested later.

For census imputations, we can tally and report the total for the estimation domains using the Census files.

For the estimation of the two remaining census components, the matching, followup and computer processing is designed to determine the enumeration status (correct or erroneous) of the

E-sample case. For the erroneous enumerations, Table 3 in Section 2.2.2 showed the different reasons for being an erroneous enumeration that needs to be determined. If we were able to determine and resolve every case, then we could just assign an indicator to each E-sample case and use that in the estimation. Since there will be some cases that we can not resolve, we will be assigning probabilities of the various outcomes to the E-sample cases. For resolved cases, we will primarily assign probabilities of '1' or '0' depending on the resolution of the case. For unresolved cases, we will assign probabilities based on a missing data model. Section 2.3.6 provides more details on the missing data adjustments and the assignments of the probabilities for component estimation.

For the component estimates that are based on the E-sample data, we will be using a two-stage ratio adjustment procedure to generate the estimates. The E sample is a representative sample of the data-defined enumerations in census housing units. The two-stage ratio adjustment is being used to reduce the variability of our estimates and ensure that the sum of our component estimates are equal to the total number of enumerations for each estimation domain (Mule and Keller upcoming 2008). Mule and Keller showed that using only one set of first-stage cells produced high standard errors for some estimates. These were seen for estimation domains that were not included in forming the first-stage ratio adjustment cells. They also showed using only a one set of first-stage cells produced component totals that did not add up to the number of enumerations for domains not included in forming the first-stage ratio adjustment cells.

The first stage will be applied to the sampling weights during the CCM estimation operations. This first stage will be done by identifying a set of cells and ratio adjusting the sum of the sampling weights so they equal the data-defined totals in those cells. These cells may be based on race/ethnicity domains, tenure, age/sex groupings or other identified variables. Depending on sample sizes and ratio adjustment factors, some of the cells may need to be collapsed together.

The second stage of ratio adjustment will be done when generating an estimate for each estimation domain. The first-stage sampling weights will be ratio adjusted to the data-defined total for the estimation domain.

When generating the component estimates laid out in the section for an estimation domain, the following approach will be used:

1. Use the Census files to determine the total census count.

2. Use the Census files to determine the number of imputations.

3. Use the Census files to determine the number of data-defined enumerations. This is also the difference of the total census count and the imputations from steps 1 and 2.

4. For the data-defined cases, we will estimate the number of those that were:

   a) correct enumerations and
   b) erroneous enumerations (overall and by type)

We will estimate the number of correct enumerations by the following formula. This is where the second-stage ratio adjustment is implemented.

$$CorrectEnumerations_C = DD_C \times \frac{\sum_{i \in C} \left(ERWEIGHT_i \times PROBCE_i \right)}{\sum_{i \in C} \left(ERWEIGHT_i \right)} \qquad (8)$$

Where $CorrectEnumerations_C$ is the estimate of correct enumerations for Estimation Domain C,
DD_C is the data-defined count for Estimation Domain C,
$ERWEIGHT_i$ is the first-stage ratio-adjusted sampling weight for E-sample case i and $PROBCE_i$ is the probability of being a correct enumeration for component estimation (different than net error estimation).

We will use the same estimator to estimate the number of erroneous enumerations. This will be done for the overall estimate and the types shown earlier. The only difference is using the corresponding probability for the desired estimate. When estimating the total number of erroneous enumerations, the overall probability of being an erroneous enumeration will be used. Based on the groupings of duplication shown in Table 3, there will be four probabilities for duplication to housing unit by geographic distance and five probabilities of duplication to group quarters by grouping of facilities. There will be one probability for erroneous enumerations for people who should not have been enumerated at all.

The estimator listed in Equation 8 generates estimates of levels for the components of Census coverage error. We realize that in addition to *levels*, estimates of *rates* will be of interest as well. These levels will be converted into rates by dividing the above estimates by the census count for the estimation domain.

Estimates of Being Enumerated in the Wrong Location

For estimates of wrong location, we will use the same estimator in Equation 8. For each estimate, the corresponding probability of being enumerated in the wrong location will be used. We will estimate the number of people enumerated in the wrong location but in the a) the right county and state, b) the wrong county in the right state and c) the wrong state. The estimation domain will be the entire nation and we will determine the probability of the case being assigned to one of the three geographic wrong locations.

2.2.5. Estimation of Omissions

Section 2.1 showed that the estimate of net error is the difference of the true population estimate and the census count in Equation 1. Net error can also be expressed as the difference between omissions and the erroneous enumerations by the following formula.

$$NetError = Omissions - ErroneousEnumerations \quad (9)$$

We can take advantage of both of these net error relationships to get an estimator of omissions. This estimator can be used for any of the estimation domains where we are generating omissions for people in housing units. Section 2.1 documented the estimation methodology for the dual system estimate and net error estimation. Section 2.2.4 documented the estimation methodology for erroneous enumerations.

$$\begin{aligned} Omissions &= NetError + ErroneousEnumerations \\ &= TruePopulationEstimate - Census + ErroneousEnumerations \end{aligned} \quad (10)$$

Where the True Population Estimate is the CCM population estimate,
Census is the census count including whole-person imputations and
Erroneous Enumerations is estimated by the methodology in the previous section.

This estimator of omissions is the starting point for the estimation. The CCM Estimation Team is researching whether a range of estimates should be given for the number of omissions based on different assumptions. The estimator would be modified to add in certain portions of the imputations.

Person Omissions by Housing Unit Inclusion Status

For people in housing units, the CCM program has been given the objective of estimating persons omissions by whether the housing unit was included or not. This is a challenge because we are being asked to estimate one quantity being missed by whether another quantity was missed. Similar to the total estimate of omissions, there is no direct estimate available so several assumptions will need to be made for us to generate this estimate in 2010.

Mule (2005) documented research done on estimation of the population in housing units enumerated in the census and the population in housing units that were missed. The document shows how using these two quantities that we can estimate the person omissions by housing unit inclusion status. We researched three possible estimators to estimate these quantities.

The first expanded on an idea that William Bell presented to the National Academy of Science panel in August 2004. The idea is that different factors may be relevant for explaining if a housing unit is enumerated in the census as compared to a person being enumerated given that the housing unit is enumerated. If this is so, then conditioning on the housing unit being

included in the census may have some benefit in the modeling and estimation. By using the housing unit matching results, we have a way of using information available for census housing units in both the E- and P-sample modeling and could examine if doing the modeling and estimation by this approach has any benefits. This allowed two dual system estimates to be calculated: a) one for the total population and b) one for the population in housing units enumerated in the census. We could estimate the persons omissions in housing units missed in the census by the difference of these two dual system estimates.

We were able to research this approach using the 2000 data because PES or Procedure C was used for the dual system estimation. For 2010 as documented in Section 2.1.4, we are using PES-B+ for the 2010 dual system estimation. For movers, this procedure mostly determines their contribution to the match rate by matching the inmovers to their given Census Day residence. The Census Day residence provided can be miles away from the sample block cluster. The matching and followup for P-sample inmovers can determine if the person matches or does not match in the search area. However, the CCM program can not always determine if the address provided by the respondent corresponds with a housing unit enumerated in that area. The CCM program can do a followup interview to determine if they have the correct Census Day address since the person now resides in the sample block cluster. The CCM program does not have the resources available to do a field followup to check other addresses located all across the country. Because of this limitation and the assumptions that would be required to use it, we have decided not to use this estimation approach for person omissions by housing unit inclusion status.

The second approach researched in Mule (2005) was to estimate these quantities based on determining the number of occupied housing that were omitted from the census. Section 3 provides the details on the net error and component estimation for housing units for the 2010 CCM. The approach estimated the occupied housing units that were omitted by adding the estimate of net error for occupied housing units and the estimate of erroneous enumerations for occupied housing units. We estimated the number of people in these missed housing units by multiplying this result by the average number of P-sample people who were in non-matched P-sample housing units. Our research using 2000 data showed that this estimate of person omissions ranged from 600,000 to 800,000 less than the estimates from the first approach. Since this approach produced this difference in population estimates, we decided not to use this approach for the 2010 estimation.

The third approach examined a “Census Plus” approach to estimate the omissions by using only the P-sample nonmatched persons in nonmatched housing units. As mentioned in the discussion of the first approach, we will not be able to determine for all cases if Census Day addresses of the inmovers were actually included or not included in the census. Because of this limitation, we have decided to not use this approach also.

We have decided to use the same methodology that was used to estimate person omissions by whether the housing unit was included or not used for the 1990 Post-Enumeration Survey (PES). Table 6 in Hogan (1993) provides results based on resolved data from housing units that had

people who did not move in or out since Census Day (nonmovers). Using these results, estimates could be generated. The estimate of person omissions by housing unit inclusion status is estimated by the proportion of resolved nonmover person nonmatches that were in housing units that either a) matched or b) did not match.

Person omissions in housing units included in the census will be estimated by:

$$PersonOmission_{HUIncluded} = Omissions \times \frac{\sum wgt_{P-sampleNonmover\ Resolved Nonmatches\ \&\ HUMatched}}{\sum wgt_{P-sampleNonmover\ Resolved Nonmatches}} \quad (11)$$

Where Omissions is the total estimate of person omissions and
wgt is the P-sample noninterview adjusted trimmed sampling weight.

Person omissions in housing units not included in the census will be estimated by:

$$PersonOmission_{HUMissed} = Omissions \times \frac{\sum wgt_{P-sampleNonmover\ Resolved Nonmatches\ \&\ HUNonmatch}}{\sum wgt_{P-sampleNonmover\ Resolved Nonmatches}} \quad (12)$$

2.3 Estimation Procedures for the Coverage of Persons in Housing Unit Estimates

The estimation methodologies described for Net Error (Section 2.1) and Components (Section 2.2) provide a general overview. This section provides more details on the estimation procedures that will be used to generate these estimates.

The estimation procedures covered in this section are:

- missing data procedures
- assigning final probabilities to sample cases for estimation
- weight trimming
- implement logistic regression models for dual system estimation
- implement the correlation bias adjustment for dual system estimation
- generate final estimates for net error
- first-stage ratio adjustment for component error
- generate final estimates component error including second-stage ratio adjustment

Missing Data to Support Estimation of the Coverage of Person in Housing Unit Estimates

2.3.1. Characteristic Imputation (E and P Sample)

Production of CCM coverage estimates requires data on age, sex, tenure (owner versus non-owner), race and Hispanic origin to classify respondents by these important demographic characteristics so they will be imputed whenever the data was not collected.

For census cases in the E sample with missing demographic characteristics, we do not have to do any special imputation procedures. We will be matching those records to the Census Edited File and obtaining the imputed characteristics from there based on the Census Edit and Imputation methodology.

For cases in the P sample with missing characteristics, we are researching to determine an imputation plan. Our current plan is impute these characteristics in a manner that is consistent with what the census is doing to the extent possible. The A.C.E. used a hot-deck procedure to assign missing tenure, race and Hispanic origin. The current plan is for the CCM to utilize the Census Edit and Imputation programs with necessary changes implemented because of the differences in how the CCM and Census data are collected. For cases with missing age, we will be imputing a single year of age. This was done for A.C.E. Revision II but not for the original A.C.E. production estimates.

For cases in either the E or P sample that are missing characteristics, the imputations will necessarily result in matched cases having the same characteristic values. While we would like the results to be as consistent as possible, we do not want to introduce a dependence in the dual system estimates by treating the matched and non-matched cases differently. We are willing to have a higher variance to keep from introducing a potential bias.

2.3.2. Noninterview Adjustment for Household Level Nonresponse

The CCM accounts for a potential bias being introduced because of households that were not interviewed. These households were not interviewed because they could not be contacted or because the interview was refused. This document focuses on P-sample noninterviews of housing units, as reported from the CCM Person Interview. These units with 'missing' interviews have sample weights associated with each, respectively. The noninterview adjustment procedure accounts for these cases by rendering a weight of zero to noninterviews and distributing the weight uniformly among interviewed cases that contain similar qualities. The distribution will be done for housing units within similar types of structures in the block cluster unless the ratio of noninteviewed to interviewed housing units is too large. If so, then the excess weighted total above the maximum allowable ratio will be distributed to other housing units in the block cluster. There is a contingency plan in place if the amount of weight distribution is too large even at the block cluster level. For the 2000 A.C.E., the maximum ratio was 2.

The CCM will use the same adjustment procedure as used for the A.C.E. except that with using PES-B+ only one noninterview adjustment for Interview Day needs to be done. The A.C.E. implemented two adjustments since they were collecting information for two points in time: one for Interview Day and a second for Census Day. In 2000, the first level of grouping was by type of basic address within a block cluster. If the adjustment did not meet certain requirements, then the additional weights were spread based on a specified hierarchy.

2.3.3. Unresolved Correct Enumeration Status for Net Error (E sample)

The CCM needs to determine if the census enumeration in the E sample was correctly enumerated in the net error search area. The CCM program has changed the imputation of enumeration status in two major ways. The first involves using a logistic regression model instead of Imputation Cell Estimation to generate the predicted probability of unresolved cases being correct enumerations.

The second involves addressing a major error in A.C.E. of overstating the number of correct enumeration. This will be done by making certain cases that have been linked to an enumeration outside of the search area as being unresolved. During clerical matching, the CCM will utilize the national computer duplicate search results to identify E-sample cases who may have a possible other Census Day residence. This additional matching may identify people who may not have fully described their living situation during the CCM interview(s)[3]. This could be an indication of possible measurement error that may introduce a bias in the logistic regression of the correct enumeration status. The missing data mechanism implemented here will attempt to minimize the potential measurement error without introducing other biases.

The E-sample cases with links to enumerations outside of the search area can be classifed into three types based on the clerical resolution of whether these two enumerations are really the same person or not:

1. Field Confirmed: When asked about possible alternate addresses where he may have been enumerated, the respondent provided the necessary information to identify the same address where the national computer matching found the case.

2. Clerical Confirmed: The respondent did not provide the necessary information to identify the same address where the national computer matching found the case. However, the clerical review of these cases is able to determine that the people are the same. One example of this is a whole family being duplicated in 2 different households. Based on the name, birth and age information provided, the clerks can determine certain cases are the same without the respondent confirmation.

[3] Either the Person Interview or the Person Followup Interview

3. Unresolved Duplication Status: The clerical review is unable to determine whether these two records linked together are the same person.

Note that it is possible for an E-sample case to have multiple duplicate links to Census records. If the multiple links are to different types (shown above) then they will be treated as having unresolved enumeration statue. Table 4 shows the possible outcomes for enumeration status based on duplication results to Census records outside the search area. The CCM program is researching the best models and covariates to use in logistic regression models to impute a probability of being a correct enumeration to unresolved cases. The possible covariates include groupings of cases based on their before followup status. Some reasons that cases are assigned to these before followup groups are a) whether they matched to PI case or not, b) whether other people in the household matched or not and c) did the case require followup. Other possible covariates include groupings based on responses to the Person Followup interview. This may involve developing separate models for cases with and without links outside the search area.

We need to assign the probability of correct enumeration for all sample cases. Resolved correct enumeration status cases will receive a probability 1. Resolved erroneous enumeration status cases will receive a probability of 0. The cases with unresolved status will have a probability assigned to them by this process.

When running the logistic regression models, the independent variables in these regressions will include matching and CCM interview information that has been identified as good discriminators of enumeration status that are only available for cases in the E sample and not available for the entire census. It will also at least include the main effects of the independent variables used in the estimation regressions shown in Section 2.1.5.

Since the overstatement of correct enumerations was a major error for the A.C.E., the CCM Estimation will closely examine and review how the imputation is being done. The details of this review are being determined. Feldpausch (2001) documents the analysis of the E-sample cases with duplicate links outside the search area for the Executive Steering Committee on A.C.E. Policy (ESCAP) II results released in October 2001.

Table 4: Possible Outcomes for Enumeration Status for Net Error Estimation Based on Duplicate Results Outside the Search Area

Duplication Results	Possible Outcomes	Assumption of Measurement Error	Description
No duplicates outside search area	Correct, Erroneous or Unresolved	No measurement error	The national matching did not identify any duplicates to suggest possible error. Depending on information collected, case will be determined to be correct, erroneous or unresolved
All duplicates outside search area have been field confirmed (No duplicates are clerically confirmed or unresolved)	Correct, Erroneous or Unresolved	No measurement error	The national matching identified duplicates at the alternate addresses provided during the interview(s). Any case resolved as correct or erroneous will be assumed to have no measurement error since the other possible address(es) were identified in interview. Case could still be unresolved based on not obtaining enough information to resolve.
All outside links that were not field confirmed have been clerically confirmed (No duplicates are unresolved)	Unresolved	Possible	The national matching identified duplicate(s) at an address that was not provided during the interview(s). Since the alternate address was not provided, we are treating these cases as unresolved to account possible measurement error.
Remaining cases where at least one outside duplicate is unresolved	Unresolved	Possible	The national matching identified potential dulicate(s) at an address that was not provided during the interview(s). . Since the alternate address was not provided, we are treating these cases as unresolved. If the two linked enumeration are the same person then we may have potential measurement error. If the two linked enumerations are not the same person, the case may have been resolved based on information collected.

2.3.4. Unresolved P-sample Inclusion Status[4] for Net Error

The CCM interview needs to determine if the person should be included in the P sample. Section 2.1.4 documents how we are using PES-B+ to identify the people in the P sample. These are residents of the sample housing unit on Interview Day. They are also outmovers from the housing unit who were residents on Census Day and who have no chance of selection based on their Interview Day residence. One example is a person who resided in the housing unit on Census Day and now resides in a group quarters on Interview Day.

For outmovers, only those who have a resolved determination that they do not reside in a housing unit on Interview Day will be included in the P sample. This decision is based on the tradeoff between having representation in our sample for this segment of the population versus introducing an error of including people who really are eligible to be included in the P sample in another housing unit.

For the remaining cases that are eligible to be included in the P sample, the CCM has implemented the two similar changes for this imputation that were described in Section 2.3.3 for the enumeration status imputation. The first is using logistic regression modeling instead of Imputation Cell Estimation.

The second involves expanding this operation to account for the potential error in overestimating the P sample. This is based on utilizing information on the cases that link[5] outside the net error search area like was done in the previous section for E-sample duplicates. The links outside the search can be classified similarly as field confirmed, clerical confirmed or unresolved links.

For nonmovers, we have determined that their Census Day residence and their Interview Day residence are the same. The finding of a link to another census enumeration outside of the search area raises the possibility that this case may not have fully reported the information about their Census Day residence. Since we have measured both the Census Day and Interview Day residence to be the same, we are assuming that we may have made an error about the Interview Day status and may be including some wrongly in the P sample. For inmovers, we are making a different assumption. We are assuming that since we found a link to another census enumeration

---

[4]"P-sample Inclusion status" is similar to the Residence status referenced in the 1990 PES and the 2000 A.C.E. Since our modification of PES-B to PES-B+ involves their residence status on both Census Day and Interview Day for some cases, we are referring to this status as "P-sample Inclusion Status" for the 2010 CCM.

[5] The word "link" is used instead of "match" in this section since the way the Census Bureau implements dual system estimation the word "match" has implications of the census enumeration being in the correct search area. We use "link" here to distinguish records linked together that are outside the search area and might be across the entire country.

that the person may not have fully reported their information and raises a possible question about whether we have captured their Interview Day status.

We will assign a probability of P-sample inclusion status for all sample cases. Resolved cases that should be included in the P sample will primarily receive a probability 1. Resolved cases that should not be included will receive a probability of 0. The cases with unresolved status need to have a probability assigned to them.

Note that it is possible for a P-sample case to have multiple links to Census records. If the multiple links are to different types (shown above) then they will be treated as having unresolved P-sample inclusion status. Table 5 shows the possible outcomes of P-sample inclusion status based on the resulting links outside the net error search area. The CCM program is researching the best models and covariates to use in logistic regression models to impute a probability of being in the P sample to unresolved cases. This may involve developing separate models for cases with and without links outside the search area.

We will use logistic regression to develop one or more models to account for cases with this unresolved status. We will implement one or more logistic regression models on the resolved cases to determine the regression coefficients. The logistic regression generates the parameter coefficients that relate the independent variables and the probability of a case being in the P sample by using iterative techniques to solve weighted likelihood equations. The weight is the P-sample weight of the case and does not include the noninterview adjustment.

We can then estimate a probability of P-sample inclusion for these unresolved cases. This will be done using the general prediction equation shown in Equation 4. Using the coefficients from the logistic regression(s) described in the previous paragraph and the values of the independent variables of the unresolved cases, we can impute the probability using the general prediction equation show in Equation 4. Since the overstatement of the P sample total was an error for the A.C.E., the CCM Estimation will closely examine and review how the imputation handled these cases. The details of this review are being determined.

Table 5: Possible Outcomes for P-sample Inclusion Status for Net Error Estimation based on Link Results Outside the Search Area

Linkage Results	Possible Outcomes	Assumption of Measurement Error	Description
No links outside search area	Included, Removed or Unresolved	No measurement error	The national matching did not identify any links to suggest possible error. Depending on information collected, case will be determined to be included in the P sample,removed from the P samples or unresolved
All links outside search area have been field confirmed (No links are clerically confirmed or unresolved)	Included, Removed or Unresolved	No measurement error	The national matching identified links at the alternate addresses provided during the interview(s). Any case resolved as included or removed s will be assumed to have no measurement error since the other possible alternate addresses were indentified in the interview Case could still be unresolved based on not obtaining enough information to resolve.
All outside links that were not field confirmed have been clerically confirmed (No links are unresolved)	Unresolved	Possible	The national matching identified link(s) at an address that was not provided during the interview(s). Since the alternate address was not provided, we are treating these cases as unresolved to account for possible measurement error.
Remainin cases where at least one outside link is unresolved	Unresolved	Possible	The national matching identified potential link(s) at an address that was not provided during the interview(s). Since the alternate address was not provided, we are treating these cases as unresolved since they may be the same person If the two linked records are the same person then we may have potential measurement error. If the two linked record are not the same person, the case may have been resolved based on information collected.

2.3.5. Unresolved Match Status for Net Error (P sample)

The CCM interview needs to determine if the P-sample person matched to a census enumeration in the net error search area. It involves accounting for P-sample cases who may have an unresolved match status or an unresolved mover status. A record may have either or both of these statuses being unresolved. This part lays out the methodology to account for both statuses and how the final match probability for net error estimation is assigned.

We will assign a probability of match status for all sample cases. Resolved match status cases will receive a probability 1. Resolved non-match status cases will receive a probability of 0. The cases with unresolved status need to have a probability assigned to them.

Since both the match status and mover status can be unresolved, we will assign an overall match probability for unresolved cases based on the following conditional probability formula.

$$\pi_{Match,j} = p_{nonmover} \times \pi_{nonmover\ Match,j} + p_{inmover} \times \pi_{inmover\ Match,j} \quad (13)$$

Where $\pi_{match,j}$ is the overall probability of being a match
$p_{nonmover}$ = probability of the case being a nonmover,
$\pi_{nonmover\ match,j}$ = probability of case matching given it is a nonmover,
$p_{inmover}$ = probability of the case being an inmover and
$\pi_{mover\ match,j}$ = probability of case matching given it is a inmover

We first need to account for unresolved mover status ($p_{nonmover}$ and p_{mover}). Resolved mover status cases will either be nonmovers or inmovers. Using the resolved cases, we will determine the weighted proportion of cases that were nonmovers and inmovers. This will be done by forming cells using some of the covariates in the logistic regression models like Before Followup groupings. This will allow the two mover probabilities to be assigned to each case: the probability of being a nonmover and the probability of being a mover. The weighted proportions will be assigned as these probabilities for the unresolved mover status cases.

We will then assign the the probabilities of a case being a match given that it is either a nonmover ($\pi_{nonmover,Match}$) case or an inmover ($\pi_{inmover,Match}$) case. When these conditional probabilities are unresolved, we will use logistic regression methods to impute them. These models and the general prediction equation will be used to determine the probability of cases matching given that the record is either a nonmover or an inmover.

For the logistic regressions, the P-sample data with resolved match status are used to determine the relationship between the dichotomous variable of being a match or nonmatch to a census enumeration and the set of independent predictor or explanatory variables. The logistic regression generates the parameter coefficients that relate the independent variables and the probability of a case being a match by using iterative techniques to solve weighted likelihood

equations. The weight used is the product of the non-interview adjusted P-sample weight assigned in Section 2.3.2. and the P-sample inclusion probability assigned in Section 2.3.4.

$$logit(match) = B_{md-m,0} + X_1 B_{md-m,1} + \ldots + X_n B_{md-m,n} \quad (14)$$

Where $Prob_{match}$ is the probability of being a match
$X_1,\ldots,X_n$ are the n independent variables and
$_{md\text{-}m,0},\ldots,\ _{md\text{-}m,n}$ are the estimated parameter coefficients of the missing data match status regression.

The predicted match rate for each unresolved match status case given that it is a nonmover is then obtained. If the case is a nonmover then the correct search area is the block cluster and ring of surrounding blocks. We first use the clerical matching information to see if they were able to match the case in the block cluster search area. If so then we will assign this conditional probability for this case equal to 1. The clerical matching information may indicate that the case was a nonmatch in the block cluster search area. If so then we will assign this conditional probability equal to 0. If the matching results are unresolved then we will assign the conditional probability by the following formula. The prediction is based on the estimated coefficients from Equation 14 and the values of the independent variables of the unresolved case. If mover status was used as a covariate in the logistic regressions and the P-sample case has an unresolved mover status then for this conditional probability assignment, the P-sample case will be treated as being a nonmover in this prediction.

$$\pi_{nonmover\ match,j} = \frac{1}{1 + e^{\left(-\beta_{md-m,0} - \sum X_{ij}\beta_{md-m,i}\right)}} \quad (15)$$

Where $\pi_{nonmover\ match,\ j}$ is the predicted match rate given the case is a nonmover for unresolved P-sample person j,
X_{ij} is the value of the *i*th independent variable for case j and
$_{md\text{-}m,i}$ is the parameter estimate of the *i*th independent variable from the missing data match logistic regression.

The predicted match rate for each unresolved P-sample case given that it is a inmover is obtained next. One difference from the conditional match probability given that the case is a nonmover is that the correct search area for inmovers needs to be identified based on interview information. The clerical results may indicate that a case links to a census enumeration outside the block cluster search area but we can not assume that it is the correct search area. For these cases the condtional match prediction will be assigned based on the following formula. The prediction is based on the estimated coefficients from Equation 14 and the values of the independent variables of the unresolved case. If mover status was used as a covariate in the logistic regressions and the

P-sample case has an unresolved mover status then for this conditional probability assignment, the P-sample case will be treated as being a inmover in this prediction.

$$\pi_{inmover\ match,j} = \frac{1}{1+e^{\left(-\beta_{md-m,0}-\sum X_{ij}\beta_{md-m,i}\right)}} \quad (16)$$

Where $\pi_{\text{inmover match, j}}$ is the predicted match rate given the case is a mover for unresolved P-sample person j,

X_{ij} is the value of the *i*th independent variable for case j and

$_{\text{md-m,i}}$ is the parameter estimate of the *i*th independent variable from the missing data match regression

After a) probabilities of the case being either a nonmover or inmover and b) the conditional probabilities of nonmovers (clerical results or Equation 15) and inmovers (Equation 16) are estimated then we can estimate the probability of the unresolved case being a match using Equation 13.

2.3.6. Unresolved Enumeration Statuses for Component Error (E sample)

Section 2.2.4. laid out the estimation of the component pieces for the census enumerations in housing units. As noted in Section 2.2, the CCM program constructed a new definition of correct or erroneous enumeration status than the definition used for dual system estimation. With this new definition, there are some cases that were determined to be erroneous enumerations for dual system estimation that will be treated as correct enumerations for component estimation. One example of the difference is that an enumeration may be erroneous for dual system estimation because it was enumerated in the wrong location but if the person was enumerated once and only once then that enumeration would be correct for national estimates of components.

Another example of a difference between the handling of cases for net error and component error are cases with Insufficient Information for DSE processing. These are treated as resolved erroneous enumerations for dual system estimation. The CCM program has expanded the matching operations to try to determine the enumeration status of these cases with minimal information for component estimation. For the matching to support component estimation, we are utilizing the scanned Census forms to see if that provides name and/or other characteristics that can help uniquely identify the Census record. We are attempting to determine enumeration status by allowing records collected during the CCM Person Interview to be matched to these cases. This matching can allow a determination of correct or erroneous to be assigned based on the result of the Person Interview for the case. For some cases, the additional matching is not able to resolve the enumeration status. Because of the lack of information collected that being usually name, these case are usually unable to go to followup to be resolved and is thus unresolved for component estimation. The imputation scheme for component estimation needs to be able to account for these unresolved cases.

As part of the estimation, the CCM Estimation needs to assign probabilities of various outcomes occurring. Because of the new definitions for components as compared to dual system estimation and having to assign probabilities of the various reasons for why a case can be erroneous, the CCM program will be implementing a new missing data adjustment of the enumeration status for component estimation. Since we are doing components for the first time, we will be using an Imputation Cell Estimation methodology to assign the probabilities of the various statuses for the unresolved cases. The A.C.E. used Imputation Cell Estimation for the missing data adjustments of unresolved statuses for the dual system estimation in 2000. This is different than the correct enumeration predictions assigned using logistic regression models for the dual system and net error estimation in Section 2.3.3.

One option would have been to use the correct enumeration probabilities assigned for dual system estimation as a starting point but that would have required the development of a) a methodology to adjust these probabilities for the new definitions and b) we would still have needed to develop an imputation methodology for the various reasons why a case may be erroneous.

Another option would have been to use logistic regression modeling. We could use the same covariates used in the logistic regression for enumeration status for dual system estimation for the logistic regressions needed for unresolved component errors. Since we are doing component estimation for the first time and we do not know if the same model should be used for each of the unresolved statuses, we are proceeding with Imputation Cell Estimation. We are more comfortable in using logistic regression methods to do the imputation of enumeration status for dual system estimation since a hierarchical logistic regression model was used in the 1990 PES (Belin et. al 1993).

The CCM Estimation staff is researching what census and CCM data can be utilized to form the best set of imputation cells for the 2010 estimation. The staff is researching if data provided during Person Followup can be utilized to provide more discriminate cells for the imputation. The weighted averages of the correct and erroneous enumerations (overall, by type using national definition and by wrong location definition) will be estimated using the resolved cases and those weighted averages will be assigned to the unresolved E-sample cases.

For component estimation, we will assign probabilities for the following unresolved statuses:

- correct enumeration nationally
- erroneous enumeration nationally

- erroneous enumeration who should not have been enumerated at all in a housing unit
- erroneous enumeration due to duplication
    - due to duplication to people in housing units by geographic distances
    - due to duplication to people in group quarters by groupings of facilities

- erroneous enumeration in the wrong location
  - right county, right state (but outside the net error search area)
  - wrong county, right state
  - wrong state

2.3.7. Final Probabilities For Net Error (E and P Samples)

Correct Enumeration for E-sample Cases

To estimate the number of correct enumerations in the census involves assigning a final probability to each E-sample case. The probability will be assigned in multiple steps. Table 6 lists how the initial correct enumeration probability is assigned for dual system estimation.

Table 6: Initial Correct Enumeration Probabilities for Dual System Estimation

Probability	
1	E-sample case is determined to be a correct enumeration.
0	E-sample case is determined to be an erroneous enumeration.
0	E-sample case has insufficient information for DSE processing
0 to 1	E-sample case has sufficient information for DSE processing but is unresolved. It is assigned a probability by logistic regression imputation (Section 2.3.3)

There is an adjustment made to the initial correct enumeration probability based on the result of duplication with persons subsampled out of the E-sample in large block clusters. The adjustment is to insure the unbiased estimation of correct enumerations in the search area. One of the E-sample people is selected during person matching to be the correct enumeration. The initial probability is multiplied by the factor below as it is not known which person was the "actual" real person. To do this, we will implement the same adjustment as was documented in Beaghen and Griffin (2000).

$$CEPROBF = CEPROBI \times \left(\frac{n_{e-sample} + 1}{n_{e-sample} + n_{non-e-sample} + 1} \right) \quad (17)$$

Where CEPROBF is the final correct enumeration probability

CEPROBI is the initial correct enumeration probability after missing data

$n_{e\text{-}sample}$ is the number of duplicate links of the E-sample case to other E-sample cases in the block cluster and

$n_{non\text{-}e\text{-}sample}$ is the number of duplicate links of the E-sample case to cases not in the E-sample cases in the block cluster.

P-Sample Inclusion and Match Probabilities for P-Sample Cases

Estimating the match rate involves assigning a probability of the P-sample case being in the P sample and a probability that the P-sample case matched to a correct enumeration in the census. Table 7 documents the various P-sample inclusion probability values assigned to P-sample cases.

Table 7: P-Sample Inclusion Probabilities for P-Sample Cases

Probability	
1	The rostered case during PI is determined to be in the P sample. The case is either a resident of the housing unit on Interview Day or is a P-sample outmover who has no chance of selection on Interview Day. P-sample outmovers are included since they are part of our target universe (Census Day residents) but have no chance of selection on Interview Day since they no longer live in a housing unit.
0	The rostered case is determined to not be in the P sample. The case did not meet one of the two criteria listed above to be in sample.
0 to 1	P-sample case with unresolved P-sample inclusion status and is assigned a probability by logistic regression imputation. P-sample cases with insufficient information for DSE processing are part of this category. (Section 2.3.4)

The match probability will be assigned as follows. Table 8 documents the various initial match probability values assigned to P-sample cases.

Table 8: Initial Match Probabilities for P-sample Cases

Probability	
1	P-sample case is determined to match to a correct enumeration.
0	P-sample case is determined to not match a correct enumeration
0 to 1	P-sample case with unresolved match status and is assigned a probability (Section 2.3.5) from the Missing Data processing. P-sample cases with insufficient information for DSE processing are part of this category.

Note: Any case on the file that was been removed from the P sample and is not part of the match rate regression will receive a final match probability of 0.

2.3.8. Final Probabilities For Component Error (E sample)

Correct Enumeration for E-sample Cases

To estimate the number of correct enumerations for components involves assigning a final probability to each E-sample case. The probability will be assigned in multiple steps similar to

those in Section 2.3.7 for net error. Table 9 lists how the initial correct enumeration probability is assigned for component estimation.

Table 9: Initial Correct Enumeration Probabilities for Component Estimation

Probability	
1	E-sample case is determined to be a correct enumeration.
0	E-sample case is determined to be an erroneous enumeration.
0 to 1	E-sample case is unresolved and assigned a probability by logistic regression imputation (Section 2.3.6)

A similar adjustment is made to the initial correct enumeration probability based on the result of duplication with persons subsampled out of the E sample in large block clusters. The adjustment is to insure the unbiased estimation of correct enumerations in the search area. One of the E-sample people is selected during person matching to be the correct enumeration. The initial probability is multiplied by the factor below as it is not known which person was the "actual" real person

$$COMPCEF = COMPCEI \times \left(\frac{n_{e-sample} + 1}{n_{e-sample} + n_{non-e-sample} + 1} \right) \quad (18)$$

Where COMPCEF is the final correct enumeration probability for components

COMPCEI is the initial correct enumeration probability for componets after missing data

$n_{e\text{-sample}}$ is the number of duplicate links of the E-sample case to other E-sample cases in the block cluster and

$n_{non\text{-}e\text{-sample}}$ is the number of duplicate links of the E-sample case to cases not in the E-sample cases in the block cluster.

*Probabilities for Erroneous Enumeration*s

Section 2.2.2 showed the types of erroneous enumerations that we will be estimating for the 2010 CCM. This part shows how the probabilities are calculated for the various types.

Duplication

To estimate duplication in the census involves assigning a probability of duplication to each E-sample case. Table 3 in Section 2.2.2 shows examples of the estimates of duplication to enumerations in housing units and group quarters. There will be four probabilities for duplication to housing unist by geographic distance and five probabilities of duplication to group quarters by grouping of facilities. Section 2.3.6 showed how we assigned a probability of being an erroneous enumeration due to duplication to unresolved cases.

An adjustment needs to be done for one of the duplication probabilities to reflect the adjustment just shown for the correct enumeration probability. In that adjustment, the correct enumeration probability was adjusted downwards by a factor for E-sample cases that had duplicate links to housing unit enumerations in the search area that were not part of the E sample. This previous adjustment *decreased* the probability of correct enumeration. To compensate, we need to *increase* the probability of erroneous enumerations. For component estimation, we will apply that difference so the change is appropriately credited as being an erroneous enumeration due to duplication to housing units within the sample block cluster search area.

We will assign the final probability of erroneous enumeration due to duplication to housing units within the sample block cluster search area (EE_DUPF1) by the following:

$$EE_DUPF1 = EE_DUPI1 + (COMPCEI - COMPCEF) \quad (19)$$

Where EE_DUPF1 is the final probability of erroneous enumeration due to duplication to housing units within the sample block cluster search area

EE_DUPI1 is the initial probability of erroneous enumeration due to duplication to housing units within the sample block cluster search area after missing data processing.

For the remaining probabilities of duplication, the final probabilities assigned will be the same as the probabilities assigned during the missing data process.

Should Not Have Been Enumerated At All

To estimate the number of erroneous enumerations due to should have not been enumerated at all involves assigning a probability to each E-sample case.

The probability of being an erroneous enumeration due to should having not been enumerated at all is assigned by the following:

= 1, if the case determined to be this type of erroneous enumeration
= 0, if the case determined to not be an erroneous enumeration for this reason (includes correct enumerations)
= probability, the case was unresolved and imputed by Imputation Cell Estimation.

Probability of Overall Erroneous Enumerations

The probability of being an overall erroneous enumeration is equal to the sum of the nine probabilities of duplication and the probability of being erroneous enumerations due to should have not been enumerated at all. It is also equal to one minus the probability of final correct enumeration.

2.3.9. Weight Trimming

The CCM Estimation Team has determined to use a similar weight trimming methodology for 2010 as was done for 2000 A.C.E. The 2000 weight trimming procedure was a modification of the 1990 procedure. Estimates of nonmatches and erroneous enumerations were generated for each block cluster and the difference of the two estimates was calculated. This cluster-level procedure only trimmed the weights if a large amount of non-offsetting errors were detected. A maximum allowable amount of non-offsetting errors was predetermined for two groups of sample block clusters: American Indian Reservations and the remaining sample. If the non-offsetting errors were greater than the maximum allowable amount, the sampling weights in the block cluster were trimmed. Using the trimmed weights, the block cluster estimate of non-offsetting errors was now equal to the maximum allowable amount. If a block cluster weight was trimmed, then the E-sample and P-sample weights were trimmed by the same magnitude. The cluster-level estimates in 2000 were based on the net error coding. We believe that this procedure will reduce the weights, if needed, of influential clusters for both the net error and component error estimation.

One change is that the weight trimming will be done after the missing data adjustments and the final probabilities have been assigned. This is a change from 2000 when the weight trimming was done before the E- and P- sample files were created using a simplified version of the missing data adjustments. Since the 2010 CCM has more time to produce estimates, we decided to move this procedure until after all of the adjustments were made so efforts would not have to be duplicated. Weight trimming adjustments will be applied to the E- and P-sample person weights.

2.3.10. Logistic Regression Models for Net Error

Section 2.1.5 gave an overview of the three logistic regression models that will be run for the dual system estimation. This section provides more details on the running of each model.

The first model involves the regression on the status of whether the census record was data-defined. This logistic regression involves using the census file after whole-person imputation has been conducted. The census processing determines our status variable of whether the record was data-defined or not. Some of the independent variables needed for the regression may need to be tallied and assigned to each census record. A software package like SAS® will be used to run the regression. Since the entire census file is used, survey weights will not be used in the logistic regression. To decrease the processing time, the data-defined and total census counts on the census file may be summarized to the cross-classification of the independent variables.

The second model involves the regression of the status of whether the data-defined census enumeration was a correct enumeration. For this regression, we will be using the results on the E-sample cases. Since we are using survey data, we want the logistic regression to reflect the

unequal probabilities of selection when solving the weighted maximum likelihood equations. The regression also needs to account for cases that received a probability between 0 and 1.

Since logistic regression needs the dependent variable to have values of '0' or '1' we create two records for each E-sample case. The two records created reflect how much each E-sample case is counted as being a correct enumeration or erroneous enumeration in the regression. When creating these new records, we will add two new variables to this new file. The first variable will be a status to be used as the dependent variable in the regression. The second variable is how much each case is counted as each status by calculating a new weight for each of these new records to use in the logistic regression.

For the first record created, the record will have a status value showing it as a correct enumeration (CESTATUS=1). The new weight will be equal to:

$$CEWEIGHT = TEWEIGHT \times CEPROBF \quad (20)$$

Where CEWEIGHT is the weight that will be used in the correct enumeration logistic regression,
TEWEIGHT is the trimmed sampling weight and
CEPROBF is the final correct enumeration probability.

For the second record created, the record will have a status value showing it is an erroneous enumeration (CESTATUS=0). The new weight will be equal to:

$$CEWEIGHT = TEWEIGHT \times (1 - CEPROBF) \quad (21)$$

The third model involves the regression of the status of whether the P-sample case matched to a correct enumeration. For this regression, we will be using the results of the P-sample cases. Since we are using survey data, we want the logistic regression to reflect the unequal probabilities of selection when solving the weighted maximum likelihood equations. This regression also needs to account for cases that received a probability between 0 and 1.

Since logistic regression needs the dependent variable to have values of '0' or '1', we create two records for each P-sample case. The two records created reflect how much each P-sample case is counted as being a match or nonmatch in the regression. When creating these new records, we will add two new variables to this new file. The first variable will be a status to be used as the dependent variable in the regression. The second variable is how much each case is counted as that status by calculating a new weight for each of these new records to use in the logistic regression. This new weight reflects the P-sample inclusion probability determined as part of the estimation processing.

For the first record created, the record will have a status value showing it is a match (MSTATUS=1). The new weight will be equal to:

$$MWEIGHT = TPNIWEIGHT \times RPROBF \times MPROBF \quad (22)$$

Where MWEIGHT is the weight that will be used in the match logistic regression,
TPNIWEIGHT is the trimmed non-interview adjusted sampling weight,
RPROBF is the P-sample inclusion probability and
MPROBF is the final match probability.

For the second record created, the record will have a status value showing it is a nonmatch (MSTATUS=0). The new weight will be equal to:

$$MWEIGHT = TPNIWEIGHT \times RPROBF \times (1 - MPROBF) \quad (23)$$

Based on the results of the three logistic regression models, the coefficients from each model will be used to assign the three necessary probabilities shown in Section 2.1.5 to each census record. These three probabilities can be used to generate a preliminary dual system estimate for each census record.

2.3.11. Correlation Bias Adjustments for Persons in Housing Units

Section 2.1.6 documented the methods used for determining the correlation bias adjustment and showed how the adjustment (c_k) is calculated to be applied to the males in the race/age groups.

The correlation bias adjustment factor, CB_j, for each census person record is equal to one of the following:

c_k if the census case is a male in the kth race/age group or

1 for all other census cases.

2.3.12. Generate Final Estimates for Net Error

We will generate the dual system estimates necessary using Equation 7 in Section 2.1.7. These estimates will be generated for the estimation domains referenced in Whitford (2008b). Using these dual system estimates, we will also estimate the net coverage of the enumeration of people in housing units by the subtraction of the census count in housing units from the dual system estimate.

In addition to the final set of requested estimates, the processing will generate the necessary files to document the estimation.

2.3.13 First-Stage Ratio Adjustment

For component estimation, we would like to control the E-sample's representation of the various populations that we will be evaluating. To reduce the variance of the component error estimates, we will ratio-adjust the weighted totals so they are equal to known census totals of data-defined enumerations. This step will implement the first stage of the ratio adjustment. The CCM estimation team is determining for which cells the sample totals should be ratio adjusted.

2.3.14. Generate Final Estimates for Component Error Including Second-Stage Ratio Adjustment

We will generate the component estimates using the estimation approaches laid out in Sections 2.2.4 and 2.2.5. These sections showed the formulas to estimate the four components of census coverage. These estimates will be generated for the estimation domains shown in Whitford (2008b). In addition to the final set of requested estimates, the processing will generate the necessary files to document the estimation.

3. Coverage Estimation of Housing Units

This section will focus on the coverage estimation for the universe of housing units. Section 3.1 provides details on the estimation of net error. Section 3.2 provides details on the measurement of component error for the 2010 CCM. Finally Section 3.3 lists the procedures and methods that will be used for net error and component error estimation.

3.1 Estimation of Net Error for Housing Units

We will estimate net error using the same approach as for persons in housing units. We will compare the estimate of the true number of housing units from the CCM program to the census count. This will be done by the following:

$$Net\,Error = True\,Population - Census \quad (24)$$

We will be implementing dual system estimation to estimate the true population of housing units. Section 2.1 provided information on the principles of dual system estimation. These principles apply to the housing unit estimates as well. The CCM program will independently list the housing units in a sample of block clusters which will be contiguous census collection blocks that average about 30 housing units. A sample of these listed housing units will be selected and those will be the P-sample housing units. We will then conduct the CCM Person interview at those housing units. While the focus is on whether the housing unit was included in the census, the CCM Person Interview results will be used to determine if the housing unit was occupied or vacant and also to identify the any of the characteristics of the householder for estimation purposes. Since we are implementing PES-B+ to handle movers for person estimation, the

occupied status and any characteristics of the householder used in estimation will be based on the CCM Interview Day. This is consistent with what was done for the 1990 PES that used PES-B.

The CCM will attempt to match the P-sample housing units to the census records. A case can only be matched if the census housing unit record is located in the correct search area. The correct search area is the CCM sample block cluster and the surrounding ring of collection blocks where the housing unit should have been counted in the census.

As part of the dual system estimation, a sample of census housing unit records, referred to as the E sample, are taken to determine how many of them were correctly counted in the right search area. The determination of the correct enumeration status for the E-sample cases is done since the general net coverage model assumes that spurious events like duplications, nonexistent cases or out-of-scope cases, such as business addresses, have been identified in both systems and accounted for in the estimation (Wolter 1986). A followup operation for both selected P-sample and E-sample cases is conducted to help resolve cases. One difference from people in housing unit net error estimation is that the address on the census files is assumed to be sufficient to identify a single housing unit. Housing unit dual system estimation does not need to implement the requirement of completeness as one of the dimensions of being correctly enumerated (Section 2.1.2).

3.1.1 Dual System Estimation Using Logistic Regression Instead of Post-stratification

Like the people in housing unit estimation, the CCM will be using logistic regression models instead of post-stratification to generate the dual system estimates. One requirement of the production logistic regression processing is that a standard statistical package like SAS® be utilized when running the logistic regressions. For more information on post-stratification for the A.C.E. housing unit estimation, see Barrett et al. (2003).

For dual system estimation of housing units, we need to run two logistic regression models to relate each of the outcomes to the chosen independent variables. Since sample data is used, the appropriate sampling weights will be used in the logistic regression. The Estimation Team is currently researching which independent variables to use in the model. When running each model, we are estimating a coefficient parameter for each variable.

The two outcomes to be modeled by logistic regression are:

1. Was an E-sample housing unit a correct enumeration in the correct location for net error estimation?

2. Did a P-sample housing unit match to a census enumeration in the correct search area?

For these two outcomes, Table 10 summarizes the successful outcome of the dependent variable, data utilized and whether weights will be utilized when running the logistic regression models. Since we are using sample data for both regressions, we will use the sampling weights in the correct enumeration and match status regressions. When running these models, the independent variables will be main effects or interactions of those effects. Our current plan is to include the same main effects in each model but not necessarily have the same interactions in each model.

Table 10: Logistic Regression Model Information for Housing Unit Dual System Estimation

Successful Outcome of the Dependent Variable	Data Source	Weights Used in Regression
Correct Enumeration	E-sample Housing Unit	Yes
Match	P-sample Housing Unit	Yes

We will use the results of the two logistic regression models to make predictions of the correct enumeration and match rate for each census case. We can make the predictions by using a) the values of the independent variables of the census case and b) the estimated parameters of the respective models shown above. The independent variables may be main effects or interactions. We will use the same general logistic regression prediction equation shown in Equation 4 in Section 2.1.5 for people in housing unit estimation. Based on the two prediction results for each census case *j*, we can then compute a dual system estimate for each case.

$$DSE_j = \frac{\pi_{ce,j}}{\pi_{m,j}} \quad (25)$$

Where DSE is the dual system estimate for census housing unit *j*.

The dual system estimates can then be computed for the national or various subdomains documented in Whitford (2008b) by :

$$DSE = \sum_{j \in C_Q} DSE_j \quad (26)$$

Where DSE is the national estimate of housing units
DSEj is the dual system estimate for each housing unit j and
C_Q are all census housing units for population Q (national, vacant, race of householder, etc.).

3.2. Component Estimation for Housing Units

For 2010, the CCM program has been given the objective of measuring the components of census coverage for housing units in addition to net error. The primary focus of the previous

coverage evaluations of the decennial censuses has focused on net error estimation. As part of the documentation of previous Housing Unit Coverage Studies, the Census Bureau released estimates of erroneous enumerations and P-sample nonmatches (Childers 1992, Barrett et al 2002) based on net error estimation.

The CCM program has determined what matching, field work and estimation would be possible to support the estimation of components of coverage for housing units in 2010. We have determined that we will only be able to do the field work to verify the status of housing units in the sample block cluster search areas for net error estimation. We will not be able to do any additional field work outside of the sample block cluster or the surrounding ring of census collection blocks. This result has implications on how to estimate components of census coverage for housing units. Ideally, we would like to followup all of the housing units anywhere in the country to determine if they were included once and only once in the census. While we do not do a personal followup outside of the sample area for the persons in housing units, we are able to utilize a national computer matching search and telephone followup to help identify if a person may have been enumerated more than once in the Census. For housing unit components, these approaches are less effective or can not be used. For component estimation, we will have the same matching and followup data as was used for dual system estimation.

For housing unit component estimation, there will be three components of census coverage.

1. Correct Enumeration

   The CCM will estimate the number of correct enumerations in the final census count. Similar to people in housing units, these are housing units that were included once and only once anywhere in the United States. If such a person or unit was included multiple times, one of the enumerations will be considered correct and the other enumerations will be erroneous.

2. Erroneous Enumerations

   The CCM will estimate the number of erroneous inclusions in the final census count. Similar to people in housing units, erroneous inclusions are enumerations that should not have been counted as a housing unit or a duplicate enumeration.

3. Omissions

   The third component is an estimate of the housing units that were omitted from the census.

For persons in housing unit estimation, we stated that using the estimated quantity of erroneous enumerations based on the stringent rules of dual system estimation can overstate the number of erroneous enumerations for components. This is less of a concern for housing unit components.

For the dual system estimation of persons, any census enumeration in the E sample without a name is classified as an erroneous enumeration because they are one type of insufficient information for dual system estimation. This automatic classification does not happen for housing unit estimation. Also, a person may move or may be associated with several different housing units where they could have been enumerated in the census. An E-sample person may have been erroneously included in the block cluster for dual system estimation but could be determined to be included once and only once in the census as a correct enumeration for components. This is different for housing units. A housing unit usually does not physically move (a trailer may move but the pad is still there) to another location.

There is an issue with the housing units that were determined to be geocoded incorrectly by the census into the sample block cluster. For dual system estimation, these housing units are treated as erroneous enumerations. For components, some of these are included once and only once in the census but we are not able to do the field verification to determine exactly where they are located. The 2000 Housing Unit Coverage Study result showed that these were a small amount (0.37 percent) of the total number of housing units and about one quarter of the size of the the E-sample housing units determined to be duplicated or not a housing unit (1.89 percent). For our estimation, we will apply a missing data model to allocate these misgeocoded census housing units as being correct or erroneous for component estimation.

3.2.1 Types of Erroneous Enumerations of Housing Units

For people in housing units, there are several reasons why a case can be an erroneous enumeration. In addition to overall estimates of erroneous enumerations, the CCM program wants to provide estimates by the reason why the case was considered erroneous. We believe that knowing the reason for the erroneous inclusions will provide more information about how to improve the 2020 Census.

For housing units, Table 11 lists the two types of erroneous enumerations that we are planning on estimating in 2010. These two types of erroneous enumerations are similar concepts to those used for people in housing unit component estimation. The 2010 CCM is not determining estimates of being included in the wrong location for housing units because we are unable to do the followup outside of the search area to identify the location of the housing units.

Table 11: Erroneous Enumerations for Housing Units

Overall Type	Description
Should not have been included in Census as a Housing Unit	This record is not a housing unit. It was either nonresidential or did not exist on Census Day. Examples of not a housing unit are group quarters, commericial or inhabitable units. Examples of nonexistent are vacant lots, demolished burned down or unable to locate
Duplication	The housing unit is a duplicate inclusion of a housing unit that was counted correctly in the census.

3.2.2. Estimation of the Components for Census Housing Units

Section 3.2 documented the two components of census coverage based on our evaluation of the housing units that were included in the census:

- Correct Enumerations
- Erroneous Enumerations

This section provides the estimation methodology to generate the estimates for an estimation domain. Whitford (2008b) documents proposed estimation domains that we are planning to evaluate for component estimation. Examples of estimation domains are the overall estimate, vacant housing units and multi-unit buildings. We are preparing to produce rate and level estimates. We have also developed our methodology to be able to produce estimates if other estimation domains are requested later.

Similar to people in housing unit component estimation, the matching, followup and computer processing is designed to determine the enumeration status (correct or erroneous) of the E-sample case. For the erroneous enumerations, Table 11 in Section 3.2.1 showed the two reasons for being an erroneous enumeration that need to be determine. If we were able to determine and resolve every case, then we could just assign an indicator to each E-sample case and use that in the estimation. Since there will be some cases that we can not resolve, we will be assigning probabilities of the various outcomes to the E-sample cases. For resolved cases, we will primarily assign probabilities of '1' or '0' depending on the resolution of the case. For unresolved cases, we will assign probabilities based on a missing data model. Section 3.3 provides more details on the missing data adjustments and the assignments of the probabilities for component estimation.

For the housing unit components that are based on the E-sample data, we will be using a two-stage ratio adjustment procedure to generate the estimates. The CCM Estimation will use a similar methodology documented for person estimation documented in Section 2.2.4. The E

sample is a representative sample of the census housing units. The two-stage ratio adjustment is being used to reduce the variability of our estimates and ensure that the sum of our component estimates are equal to the total number of housing units for each estimation domain.

The first stage will be applied to the sampling weights during the CCM estimation operations. This first stage will be done by identifying a set of cells and ratio adjusting the sum of the sampling weights so they equal the housing unit totals in those cells. These cells may be based on region, occupancy status or other identified variables. Depending on sample sizes and ratio adjustment factors, some of the cells may need to be collapsed together.

The second stage will be done when generating an estimate for each estimation domain. The first-stage sampling weights will be ratio adjusted to the census housing unit total for the estimation domain.

When generating estimates for an estimation domain, the following steps will be used. Since imputations are not one of the components for housing unit estimation, steps from Section 2.2.4 have been removed.

1. Use the Census files to determine the total census housing unit count.

2. For the housing unit estimation, we will estimate the number of those that were

   a) correct enumerations and
   b) erroneous enumerations

We will estimate the number of correct enumerations by the following formula. This is where the second-stage ratio adjustment is implemented.

$$CorrectEnumerations_C = CENHU_C \times \frac{\sum_{i \in C} \left(ERHUWEIGHT_i \times PROBCE_i \right)}{\sum_{i \in C} \left(ERHUWEIGHT_i \right)} \quad (27)$$

Where $CorrectEnumerations_C$ is the estimate of correct enumerations for Estimation Domain C,
$CENHU_C$ is the census housing unit count for Estimation Domain C,
$ERHUWEIGHT_i$ is the first-stage ratio-adjusted sampling weight for E-sample housing unit case i and
$PROBCE_i$ is the probability of being a correct enumeration for component estimation (different than net error estimation).

We will use the same estimator to estimate the number of erroneous enumerations for both overall and by type. The only difference is using the probability of being erroneous for the estimate. The estimates of levels generated in Equation 27 can be converted into rates by dividing the above estimates by the census count for the estimation domain.

3.2.3. Estimation of Housing Unit Omissions

We will use the same estimator for housing unit omissions that was used for people in housing unit omissions in Section 2.2.3. Section 3.1 documented the estimation methodology for the dual system estimate and net error estimation for housing units. Section 3.2.3 documented the estimation methodology for erroneous enumerations.

$$\begin{aligned} Omissions &= NetError + ErroneousEnumerations \\ &= TruePopulationEstimate - Census + ErroneousEnumerations \end{aligned} \quad (28)$$

Where the True Population Estimate is the CCM population estimate,
Census is the census count of housing units,
and Erroneous Enumerations is estimated by the methodology in the previous section.

3.3 Estimation Procedures for the Coverage of Housing Unit Estimates

The estimation methodologies described for net error (Section 3.1) and components (Section 3.2) provide a general overview. This section provides more details on the estimation procedures that will be used to generate these estimates.

The estimation procedures covered in this section are:

- missing data procedures
- assigning final probabilities to sample cases for estimation
- weight trimming
- implement logistic regression models for dual system estimation
- generate final estimates for net error
- first-stage ratio adjustment for component errors
- generate final estimates for component error including second-stage ratio adjustment

Missing Data to Support Estimation of the Coverage of Housing Unit Estimates

3.3.1. Characteristic Imputation (E and P Samples)

Production of CCM coverage estimates requires data on occupancy status, tenure (owner versus non-owner), race/ethnicity domain of the householder to classify housing units by these important characteristics so they will be imputed whenever the data was not collected. For census cases in the E sample with missing housing unit characteristics, we do not have to do any special imputation procedures. We will be using the Census edited file and obtaining the imputed characteristics from there.

For housing units in the P sample with missing characteristics, we are researching an imputation plan. The imputation of tenure for occupied housing units will be the same process done for person estimation. The remaining imputation for housing unit estimation is new for 2010 and different than the 2000 A.C.E. The A.C.E. implemented a noninterview adjustment for sample cases when the characteristic of the race/ethnicity domain of the householder was not able to be determined because the person interview was not completed. For housing unit estimation, we can still match and followup the housing unit without this missing characteristic so for 2010 we do not want to implement a noninterview adjustment and essentially throw sample away. For 2010, we will handle this missing characteristic like the other missing characteristics for person and housing unit estimation and it will be imputed by a method similar to the Census Edit and Imputation system.

3.3.2. Unresolved Correct Enumeration Status for Net Error (E sample)

The CCM needs to determine if the census housing unit in the E sample was correctly enumerated in the net error search area. We need to assign the probability of correct enumeration for all sample cases. Resolved correct enumeration status cases will receive a probability of 1. Resolved erroneous enumeration status cases will receive a probability of 0. The cases with unresolved status need to have a probability assigned to them.

We will use logistic regression to develop one or more models to account for cases with this unresolved status. We will run one (or more if necessary) logistic regression models on the resolved cases to determine regression coefficients.

The logistic regression generates the parameter coefficients that relate the independent variables and the probability of the case being a correct enumeration by using iterative techniques to solve weighted likelihood equations. The independent variables in these regressions can include more independent variables used in the regressions shown in Section 3.1.1. These models can include relevant predictors like Housing Unit Followup responses that are only available for the cases in the E sample and not available for the entire census. Using these coefficients and the data of the unresolved cases we can then estimate a probability of correct enumeration for these unresolved cases.

3.3.3. Unresolved Housing Unit Status for Net Error (P sample)

The CCM Housing Unit matching operation needs to determine if a P sample housing unit was a housing unit on Census Day. For some units in sample, this will not be able to be determined and they will be unresolved. We will assign a probability of housing unit status for all sample cases. Resolved housing units will receive a probability of 1. Resolved cases determined not to be housing units will receive a probability of 0. The cases with unresolved status need to have a probability assigned to them.

We will use logistic regression to develop one or more models to account for cases with this unresolved status. We will implement one or more logistic regression models on the resolved cases to determine the regression coefficients. The logistic regression generates the parameter coefficients that relate the independent variables and the probability of a case being a match by using iterative techniques to solve weighted likelihood equations. The weight is the P-sample weight, the inverse of the probability of selection. Using these coefficients and the data of the unresolved cases we can then estimate a probability of housing unit for these unresolved cases.

3.3.4. Unresolved Match Status for Net Error (P sample)

The CCM matching operation needs to determine if the housing unit matched a census enumeration in the net error search area. We will assign a probability of match status for all sample cases. Resolved match status cases will receive a probability of 1. Resolved nonmatch status cases will receive a probability of 0. The cases with unresolved status need to have a probability assigned to them.

We will use logistic regression to develop one or more models to account for cases with this unresolved status. We will implement one (or more if necessary)logistic regression models on the resolved cases to determine the regression coefficients. The logistic regression generates the parameter coefficients that relate the independent variables and the probability of a case being a match by using iterative techniques to solve weighted likelihood equations. The weight is the P-sample weight. The same weight is used for the housing unit and match status regressions since there is no noninterview adjustment for housing unit estimation. Using these coefficients and the data of the unresolved cases we can then estimate a probability of match for these unresolved cases.

3.3.5. Unresolved Enumeration Statuses for Component Error (E sample)

Section 3.2.3 laid out the estimation of the component pieces for the census enumerations in housing units. As noted in Section 3.2, the CCM program will be using a different definition of correct or erroneous enumeration status than the definition used for dual system estimation. With this new definition, there are some cases that were determined to be erroneous

enumerations for dual system estimation that will be treated as unresolved enumerations for component estimation.

One example of the difference is that an enumeration may be erroneous for dual system estimation because it was a geocoding error. For component estimation, a housing unit is correct if it was included once and only once in the census. These geocoding errors have been erroneously assigned to the sample block cluster. For our estimation, the missing data model will assign probabilities to these misgeocoded census housing units to reflect their being correct or erroneous for component estimation. The missing data procedure will assign probabilities of erroneous overall and also by the two types listed in Table 11 in Section 3.2.

3.3.6. Final Probabilities For Net Error

Correct Enumeration for E-sample Cases

To estimate the number of correct enumerations in the census involves assigning a final probability to each E-sample case. The probability will be assigned in multiple steps. Table 12 lists how the initial correct enumeration probability is assigned for dual system estimation.

Table 12: Initial Correct Enumeration Probabilities for Dual System Estimation

Probability	
1	E-sample case is determined to be a correct enumeration.
0	E-sample case is determined to be an erroneous enumeration.
0 to 1	E-sample case has unresolved enumeration status and is assigned a probability by logistic regression imputation

The final correct enumeration probability is assigned by adjusting the initial correct enumeration probability by a factor to insure the unbiased estimation of correct enumerations in the search area. To do this, we will implement the same adjustment as was documented in Beaghen and Griffin (2000).

$$CEPROBF = CEPROBI \times \left(\frac{nhu_{e-sample} + 1}{nhu_{e-sample} + nhu_{non-e-sample} + 1} \right) \quad (29)$$

Where CEPROBF is the final correct enumeration probability,
CEPROBI is the initial correct enumeration probability,
$nhu_{e\text{-}sample}$ is the number of duplicate housing unit links of the E-sample housing unit to other E-sample housing units cases in the block cluster and
$nhu_{non\text{-}e\text{-}sample}$ is the number of duplicate housing unit links of the E-sample housing unit to housing unit cases not in the E-sample cases in the block cluster.

3.3.7. Final Probabilities For Component Error (E Sample)

Probability of Correct Enumeration

A similar adjustment is made to the initial correct enumeration probability based on the result of duplication with housing units subsampled out of the E sample in large block clusters. The adjustment is to insure the unbiased estimation of correct enumerations in the search area. One of the E-sample housing units is selected during matching to be the correct enumeration. The initial probability is multiplied by the factor below as it is not known which housing unit was the "actual" housing unit.

$$COMPCEF = COMPCEI \times \left(\frac{nhu_{e-sample} + 1}{nhu_{e-sample} + nhu_{non-e-sample} + 1} \right) \qquad (29)$$

Where COMPCEF is the final correct enumeration probability for components
COMPCEI is the initial correct enumeration probability for components,
$nhu_{e\text{-}sample}$ is the number of duplicate housing unit links of the E-sample housing unit to other E-sample housing units cases in the block cluster and
$nhu_{non\text{-}e\text{-}sample}$ is the number of duplicate housing unit links of the E-sample housing unit to housing unit cases not in the E-sample cases in the block cluster.

*Probabilities for Erroneous Enumeration*s

Section 3.2.4 showed the types of erroneous enumerations that we will be estimating for the 2010 CCM. This part shows how the probabilities are calculated for the various types.

Duplication

An adjustment needs to be done to the probability to reflect the adjustment just shown for the correct enumeration probability. In that adjustment, the correct enumeration probability was adjusted downwards by a factor for E-sample cases that had duplicate links to housing unit enumerations in the search area that were not part of the E sample. This previous adjustment *decreased* the probability of correct enumeration. To compensate, we need to *increase* the probability of erroneous enumerations. For component estimation, we will apply that difference so the change is appropriately credited as being an erroneous enumeration due to duplication to housing units within the sample block cluster search area.

We will assign the final probability of erroneous enumeration due to duplication by the following:

$$EE_DUPF = EE_DUPI + \left(COMPCEI - COMPCEF \right) \qquad (30)$$

Where EE_DUPF is the final probability of duplication due to duplication
EE_DUPI is the probability of duplication assigned after missing data processing.

Should Not Have Been Enumerated At All

To estimate the number of the census housing unit records that should not have been enumerated at all involves assigning a probability to each E sample case. A probability will be assigned to each E-sample case by the following:

= 1, if the case determined to be this type of erroneous enumeration
= 0, if the case determined to not an erroneous enumeration for this reason (includes correct enumerations)
= probability, the case was unresolved and imputed by Imputation Cell Estimation (Section 3.3.5).

3.3.8. Weight Trimming

The decision is to use a similar weight trimming methodology for housing unit estimates that was used for the 2000 A.C.E. As in 2000, there is a separate trimming procedure for housing unit estimation. The 2000 housing unit weight trimming procedure was a modification of the 1990 person procedure. Estimates of nonmatches and erroneous enumerations were generated for each block cluster and the difference of these two estimates was calculated. This cluster-level procedure only trimmed the weights if a large amount of non-offsetting errors were detected. A maximum allowable amount of non-offsetting errors was predetermined for two groups of sample block clusters: American Indian Reservations and the remaining sample. If the non-offsetting errors were greater than the maximum allowable amount, the sampling weights in the block cluster were trimmed. Using the trimmed weights, the block cluster estimate of non-offsetting errors was now equal to the maximum allowable amount. If a block cluster was trimmed, then the E-sample and P-sample weights were trimmed by the same magnitude. The cluster-level estimates in 2000 were based on the net error coding. The timing of the housing unit weight trimming being done after the missing data adjustments and the final probabilities have been assigned is the same as was done for the A.C.E Weight trimming adjustments would be applied to the E- and P-sample housing unit weights.

3.3.9. Logistic Regression Models for Net Error

Section 3.1.1 gave an overview of the two logistic regression models that will be run for the dual system estimation of housing units. This section provides more details on the running of each model.

The first model involves the regression of the status of whether the housing unit was a correct enumeration. For this regression, we will be using the results on the E-sample cases. Since we are using survey data, we want the logistic regression to reflect the final weights when solving

the weighted maximum likelihood equations. The regression also needs to account for cases that received a probability between 0 and 1.

Since logistic regression needs the dependent variable to have values of ‘0’ or ‘1’, we create two records for each E-sample case. The two records created reflect how much each E-sample case is counted as being a correct enumeration or erroneous enumeration in the regression. When creating these new records, we will add two new variables to this new file. The first variable will be a status to be used as the dependent variable in the regression. The second variable is how much each case is counted as each status and determined by calculating a new weight for each of these new records to use in the logistic regression.

For the first record created, the record will have a status value showing it is a correct enumeration (CESTATUS=1). The new weight will be equal to:

$$CEWEIGHT = TEWEIGHT \times CEPROBF \quad (32)$$

Where CEWEIGHT is the weight that will be used in the correct enumeration logistic regression,
TEWEIGHT is the trimmed sampling weight and
CEPROBF is the final correct enumeration probability.

For the second record created, the record will have a status value showing it is an erroneous enumeration (CESTATUS=0). The new weight will be equal to:

$$CEWEIGHT = TEWEIGHT \times (1 - CEPROBF) \quad (33)$$

The second model involves the regression of the status of whether the P-sample case matched to a correct enumeration. For this regression, we will be using the results of the P-sample cases. Since we are using survey data, we want the logistic regression to reflect the final weight when solving the weighted maximum likelihood equations. This regression also needs to account for cases that received a probability between 0 and 1.

Since logistic regression needs the dependent variable to have values of ‘0' or ‘1', we create two records for each P-sample case. The two records created reflect how much each P-sample case is counted as being a match or nonmatch in the regression. When creating these new records, we will add two new variables to this new file. The first variable will be a status to be used as the dependent variable in the regression. The second variable is how much each case is counted as each status and is determined by calculating a new weight for each of these new records to use in the logistic regression.

For the first record created, the record will have a status value showing it is a match (MSTATUS=1). The new weight will be equal to:

$$MWEIGHT = TPWEIGHT \times HUPROB \times MPROBF \quad (34)$$

Where MWEIGHT is the weight that will be used in the match logistic regression,
TPWEIGHT is the trimmed sampling weight
HUPROB is the housing unit status probability and
MPROBF is the final correct enumeration probability.

For the second record created, the record will have a status value showing it is a nonmatch (MSTATUS=0). The new weight will be equal to:

$$MWEIGHT = TPNIWEIGHT \times HUPROB \times (1 - MPROBF) \quad (35)$$

Based on the results of the two logistic regression models, the coefficients from each model will be used to assign the necessary probabilities shown in Section 3.1.1 to each census record. These three probabilities can be used to generate a preliminary dual system estimate for each census record.

3.3.10. Final Estimates for Net Error

We will generate the dual system estimates necessary using Equation 26 in Section 3.1.1. These estimates will be generated for the estimation domains listed in Whitford (2008b). Using these dual system estimates, we will also estimate the net coverage of the enumeration of people in housing units by the subtraction of the census count in housing units from the dual system estimate. In addition to the final set of requested estimates, the processing will generate the necessary files to document the estimation.

3.3.11 First-Stage Ratio Adjustment

For component estimation, we would like to control the E-sample's representation of the various populations that we will be evaluating. To reduce the variance of the component error estimates, we will ratio-adjust the weighted totals in two steps so they are equal to known census totals of housing units. This step will implement the first stage of the ratio adjustment. The CCM estimation team is determining for which cells the sample totals should be ratio adjusted. The cells used for the housing unit adjustment may be different than the cells used for the person adjustment.

3.3.12. Final Estimates for Component Error Including Second-Stage Ratio Adjustment

We will generate the component estimates using the approaches laid out in Sections 3.2.2 and 3.2.3. These sections showed the formulas to estimate the three components of census coverage.

These estimates will be generated for the estimation domains listed in Whitford (2008b). In addition to the final set of requested estimates, the processing will generate the necessary files to document the component estimation.

References

Barrett, D., Beaghen, M., Smith, D., and Burchum, J. (2001), "ESCAP II: Housing Unit Coverage Study," Executive Steering Committee for A.C.E. Policy II, Report 20.

Barron, W., (2001, July 13), "Statement by William G. Barron, Jr. on the Current Status of Results of Census 2000 Accuracy and Coverage Evaluation Survey," U.S. Census Bureau.

Beaghen, M. And Griffin, R. (2001), "Accuracy and Coverage Evaluation Survey's Probability of Correct Enumeration for Duplicated Census Enumerations," DSSD Census 2000 Procedures and Operations Memorandum Series Q-58.

Belin, T., Diffendal, G., Mack, S., Rubin, D., Schafer, J. and Zaslavsky A. (1993), "Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation." Journal of the American Statistical Association, 88, 1149-1166.

Bell, W. (1999), "Accuracy and Coverage Evaluation Survey: Combining Logistic Regression DSEs (Target Model) with 1990 Demographic Analysis Results," DSSD Census 2000 Procedures and Operations Memorandum Series Q-12.

Chandrasekar, C. and Deming, W. E. (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration". Journal of the American Statistical Association, 44, 101-115.

Childers, D., (1992), "The 1990 Housing Unit Coverage Study," *Proceedings.* Survey Research Methods Section. American Statistical Association, Alexandria, VA.

Feldpausch, R. (2001), "ESCAP II: Census Person Duplication and the Corresponding A.C.E. Enumeration Status," Executive Steering Committee for A.C.E. Policy II, report 6.

Griffin, R. (2000), "Accuracy and Coverage Evaluation Survey: Definition and Explanation of Correlation and Related Heterogeneity Bias" DSSD Census 2000 Procedures and Operations Memorandum Series Q-35.

Griffin (2005), "Net Error Estimation for the 2010 Census", DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-1.

Haberman, S. J., Jiang. W., and Spencer B. D. (1998), " Activity 7: Develop Methodology for Evaluating Model-Based Estimates of the Population size for States Final Report", prepared by NORC for the Bureau of the Census under contract no. 50-YABC-2-66023.

Hogan, H. (1992), "The 1990 Post-Enumeration Survey: An overview." The American Statistician, 46, 261-269.

Hogan, H. (1993), "1990 Post-Enumeration Survey: Operation and Results," Journal of the American Statistical Association.

Hogan, H. (2003), "The Accuracy and Coverage Evaluation: Theory and Design." Survey Methodology, 29, 129-138.

Judson, D.H., J. Gregory Robinson, Kirsten West, Antonio Bruce, Bashiruddin Ahmed, and Frederick W. Hollmann, A (2008), "Demographic Analysis: Options in 2008 and Beyond", paper presented at the Joint Statistical Meetings, Denver, Colorado, August 2008.

Kostanich, D. (2003), "A.C.E. Revision II: Design and Methodology," DSSD A.C.E. Revision II Memorandum Series, #PP-30, Washington, D.C.

Marks, E.S. (1979), The Role of Dual System Estimation in Census Evaluation. In Recent Developments in PGE, (K. Krotki). University of Alberta Press. 156-188.

Moldoff, M. (2008), "The Design of the Coverage Measurement Program for the 2010 Census" DSSD 2010 Census Coverage Measurement Memorandum Series #2010-B-7.

Mule, T. (2005), "Preliminary Research on Separate Estimation of the Population in Enumerated and Non-Enumerated Housing Units" DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-6.

Mule, T. and Keller A. (Upcoming 2008), "Pseudo-Estimation of Components for the 2006 Census Coverage Measurement" DSSD 2010 Census Coverage Measurement Memorandum Series.

Mulry, M. H., Spencer, B. D., Mule, T., Nguyen, N. (2008), "Direct Estimates As a Diagnostic for Dual System Estimators Based on Logistic Regression". JSM *Proceedings.* Survey Research Methods Section. American Statistical Association, Alexandria, VA.

Petersen, C.G.J. (1896), The Yearly Immigration of Young Plaice into the Limfjord from the German Sea. Report of the Danish Biological Station. 6, 1-48.

Shores, R. (2003), "A.C.E. Revision II: Adjustment for Correlation Bias," DSSD A.C.E. Revision II Memorandum Series, #PP-53, Washington, D.C.

Singh, R. (2003), "2010 Census Coverage Measurement - Goals and Objectives (Executive Brief)," DSSD 2010 Census Coverage Measurement Memorandum Series.#A-1, Washington, D.C.

Singh, R. (2005), "2010 Census Coverage Measurement - Updated Plans," DSSD 2010 Census Coverage Measurement Memorandum Series #A-6, Washington, D.C.

U.S. Census Bureau (1985), Evaluating Census of Population and Housing, Statistical Training Document, ISP-TR-5, Washington, D.C.

U.S. Census Bureau (2004), Accuracy and Coverage Evaluation of Census 2000: Design and Methodology, Washington, D.C.

Whitford, D. (2008a), "Overview of the 2010 Census Coverage Measurement Program" DSSD 2010 Census Coverage Measurement Memorandum Series #A-19, Washington, D.C.

Whitford, D. (2008b), "Proposed Census Coverage Measurement Estimates for Net and Component Error" DSSD 2010 Census Coverage Measurement Memorandum Series #A-23, Washington, D.C.

Wolter, K.M., (1986), "Some Coverage Error Models for Census Data" Journal of the American Statistical Association, 81, 338-353.

Appendix A: Net Coverage Estimation for the 2000 Accuracy and Coverage Evaluation

Post-stratification Approach for Dual System Estimation

The post-stratification approach in the original A.C.E. estimated the following quantities for each of the 416 post-strata to get the dual system estimate. Each census, E-sample and P-sample case is assigned to one and only one post-stratum and is used to calculate the needed parts.

$$DSE_I = C \times r_{dd} \times \frac{r_{ce}}{r_m} \qquad (2)$$

Where C is the census count,
r_{dd} is the data-defined rate (excluding reinstates) of the census,
r_{ce} is the correct enumeration rate from the E sample and
r_m is the match rate from the P sample.

Another way to write the post-stratification estimator is by the following summation of census cases:

$$DSE_I = \sum_{j \in I} r_{dd(j)} \times \frac{r_{ce(j)}}{r_{m(j)}} \qquad (3)$$

Where $r_{dd(j)}$ is the data-defined rate for person j who is in post-stratum I,
$r_{ce(j)}$ is the correct enumeration rate for person j who is in post-stratum I,
And $r_{m(j)}$ is the data-defined rate for person j who is in post-stratum I.

Handling Movers and the Post-stratification Estimators for A.C.E.

For person estimates, dual system estimation needs to account for the population that moves between Census Day and the day of the coverage measurement interview. People who have moved into the housing unit since Census Day are inmovers and people who have moved out of the housing unit since Census Day are outmovers. The CCM in 2010 will be using a version of the "PES-B" methodology and just for the inmovers. It will use the results of matching the inmovers to their Census Day address to account for movers. The A.C.E. obtained information on both groups of people in 2000 so that a "PES-C" methodology could be used. PES-C uses the results of matching the outmovers in the cluster search area to determine the match rate of the mover population and the number of inmovers to estimate the total number of movers. If not enough outmovers were present then the PES-A methodology using just outmovers was used instead.

PES-C Estimator

The following is the PES-C estimator. This uses both inmovers and outmovers to account for the mover population in the P sample. The outmovers are used to estimate the match rate of the mover population and the inmovers are used to estimate the total number. This was used for any post-stratum that had 10 or more outmover cases.

$$DSE = (C^* - II^*) \times \frac{CE}{N_e} \times \frac{N_n + N_i}{M_n + \frac{M_o}{N_o} \times N_i}$$

Where C^* is the census count,
II^* is the insufficient information or reinstated data-defined cases,
CE is the estimate of correct enumerations from the E sample,
N_e is the estimate of enumerations from the E sample,
M_n is the estimate of nonmover matches from the P sample,
M_o is the estimate of outmover matches from the P sample,
N_n is the estimate of nonmovers from the P sample
N_o is the estimate of outmovers from the P sample and
N_i is the estimate of inmovers from the P sample.

PES-A Estimator

The PES-A estimator uses only outmovers to account for the mover population in the P sample. The outmovers are used to estimate the match rate of the mover population and are used to estimate the total number. This was used for any post-stratum that had fewer than 10 outmover cases.

$$DSE = (C^* - II^*) \times \frac{CE}{N_e} \times \frac{N_n + N_0}{M_n + M_o}$$

Where C^* is the census count,
II^* is the insufficient information or reinstated data-defined cases,
CE is the estimate of correct enumerations from the E sample,
N_e is the estimate of enumerations from the E sample,
M_n is the estimate of nonmover matches from the P sample,
M_o is the estimate of outmover matches from the P sample,
N_n is the estimate of nonmovers from the P sample
N_o is the estimate of outmovers from the P sample.

Appendix B Logistic Regression Modeling

The DSSD is requiring that the logistic regression models being run using a statistical software package like SAS. The reason is that logistic regression with survey data involves estimating the regression coefficient parameters by a weighted maximum-likelihood approach using either a Newton-Raphson approach or the Fisher scoring algorithm. Either of these approaches could be specified and programmed. Each would require a programming of the algorithm and also involves the inverting of the data matrix. This matrix will have at least one row for each case in sample (700,000) and then have a column for each covariate in the model which could be in the hundreds. A separate new programming of the algorithm would be very resource intensive. The DSSD is most familiar with using SAS software to conduct logisitic regression so that will be used in stating what statistical software packages can do.

The logistic regression models will have the form using correct enumeration status as an example. When using sample data, the appropriate sampling weight will be used.

$$logit(Prob_{ce}) = B_{ce,0} + X_{ce,1}B_{ce,1} + \ldots + X_{ce,n}B_{ce,n} + e$$

Where $Prob_{ce}$ is the probability of being a correct enumeration

$X_{ce,1},\ldots,X_{ce,n}$ are the n independent variables for the correct enumeration status regression and

$_{ce,0},\ldots,\ _{ce,n}$ are the parameter coefficients estimates of the n independent variables for the correct enumeration status regression.

As shown in Equation 4 in Section 2.1.5 shows the general prediction equation. Using the appropriate coefficients ( )and the covariate (X) values, we can assign the predicted probability to the case.

SAS's Logistic regression procedure (PROC LOGISTIC) can run a logistic regression model of our binary status outcomes against our specified covariates for our sample cases. It can do the following:

- use the survey weights when running the model
- generates automatic output listing that documents the regression
- output the estimated coefficieints to a dataset for later use
- estimated predicted values of input dataset can be computed and saved on an output version.
- Allows the "Scoring of a New Dataset"
    - Creates special file when running each model that can be saved.
    - Apply a previously fitted model using this special file to a new data set in order to compute the probabilities of each response category given the values of the explanatory variables in each observation.