Total Error Model for Census 2000:
How Components of Error Can Be Estimated from the Bureau's Planned Evaluation Studies

Final Report

Bruce D. Spencer

May 11, 2000; rev. March 20, 2002

Prepared by Abt Associates Inc. and Spencer Statistics, Inc.
for the Bureau of the Census

Task Number 46-YABC-7-00001, Activity 12.18

Contract Number 50-YABC-7-66020

A. Introduction

This report discusses the definition and estimation of components of error needed for total error model for Census 2000. Although the total error model has been described elsewhere, to make this report relatively self-contained we will repeat the description of the total error model and its components. We proceed in two steps. First (Section B) we take an abstract view of the dual system estimator (DSE). We begin by introducing a probability model for the census and then we describe in general terms the use of poststratification for estimating probabilities of individuals being missed and enumerations being erroneous. We then discuss the overall decomposition of error into random sampling error, ratio estimator bias, measurement error bias, model bias, contamination bias, and bias due to uncountable persons. Model bias is further examined and shown to consist of correlation bias and synthetic estimation bias. Next (Section C) we examine the DSE and its components in great detail, with the most discussion devoted to the components of measurement error. In Section C we will briefly discuss the estimation of the components of error.

B. An Abstract View of the DSE

1. A Probability Model for the Census[1]

Consider a set S_0 with N_0 potential census enumerations for the population. Let J be a nominal variable defined on S_0 with possible values EE (erroneous enumeration), C (correct enumeration), and U (unenumerated). Let S_T be the set of s in S_0 such that J(s) equals C or U, i.e., the "true population". For x equal to EE, C, or U, let $\delta_x(J)$ be the variable with value 1 if J = x and value 0 if $J \neq x$, and let $\delta_x(J(s))$ be the value of $\delta_x(J)$ for s in S. The correct population total is N given by

$$N = \sum_{s \in S_0} [1 - \delta_{EE}(J(s))],$$

the number of members in S_T. Let S_E be the set of s in S_0 such that J(s) is EE or C, so that S_E consists of all census enumerations and is of size

$$N_E = \sum_{s \in S_0} [1 - \delta_U(J(s))].$$

It is convenient to consider probabilities of enumeration statuses EE, C, and U. Thus, let J(s) be a random variable for s in S_0, where J(s) is x with probability $p_x(s)$ for x equal EE, C, or U. Define the adjustment factor (or the *true* adjustment factor) by

[1]This section uses a model developed in Haberman, Jiang, and Spencer (1998).

$$A(s) = \begin{cases} \dfrac{p_C(s) + p_U(s)}{p_C(s) + p_{EE}(s)} = \dfrac{1 - p_{EE}(s)}{1 - p_U(s)}, & p_U(s) < 1 \\ 0, & p_U(s) = 1. \end{cases}$$

(The value of $A(s)$ when $p_U(s) = 1$ does not matter because it cannot be applied.) If $p_x(s)$ are known for each s in S_E, then N may be approximated by the random variable

$$\sum_{s \in S_E} A(s).$$

If $p_U(s) < 1$ for each s, this approximation has expectation N; otherwise, the expectation equals $N - N_U$, with N_U the number of members s for which $p_U(s) = 1$.

2. Poststratification

In practice, estimates of factors A(s) must be based on some modeling. The modeling is unavoidable because there are at most two opportunities for an individual to be enumerated, the census and the P sample. Thus, individuals with similar characteristics are modeled as having similar probabilities of being missed, and enumerations with similar recorded characteristics are modeled as having similar probabilities of being erroneous. Even for a sampled block, the number of persons will typically be small enough that pooling data across blocks will improve the accuracy of the estimates of A(s). Pooling data across blocks clearly must be used to estimate A(s) for individuals in nonsampled blocks.

The Census Bureau will estimate the true adjustment factors A(s) by assuming they are constant within poststrata. P-sample persons are grouped into poststrata and within each poststratum the sample-weighted fraction of P-sample persons who were missed in the census is computed[2]; that fraction is used for $\hat{p}_U(s)$ if person s belongs to the poststratum. E-sample enumerations are grouped into poststrata and within each poststratum the sample-weighted fraction of E-sample enumerations that were erroneous is computed[3]; that fraction is used for $\hat{p}_{EE}(s)$ if enumeration s belonged to the poststratum. The poststratified DSE can be written in the form

$$\hat{N} = \sum_{s \in S_E} \hat{A}(s)$$

---

[2]Cases with missing data are assigned imputed probabilities of being missed in the census.

[3]Cases with missing data are assigned imputed probabilities of being erroneous. The fraction of enumerations that are not erroneous is ratio-adjusted by the fraction of census enumerations that are data defined; see Section C.1.

with

$$\hat{A}(s) = \frac{1 - \hat{p}_{EE}(s)}{1 - \hat{p}_U(s)}.$$

(It is more efficient to group the enumerations by poststratum when computing $\hat{N}$, but the results are algebraically the same. The Census Bureau uses the same poststrata for $\hat{p}_U(s)$ and $\hat{p}_{EE}(s)$, but that is not necessary.)

3. Overview of Error Components

The estimates $\hat{A}$ are based on a sample of housing units in a sample of blocks and as such are subject to two kinds of sampling error, random error and bias. Random sampling error arises because only a sample are used to develop the estimates. The estimated factors $\hat{A}(s)$ involve products and ratios of random variables, and as such are subject to a technical bias known as "ratio estimator bias"; Cochran (1977, p. 161). Let E(.) denote expected value; note that expectation is taken with respect to sampling as well as probabilistic aspects of behavior in the census and A.C.E.. Let $\hat{A}_1(s)$ denote the value of $\hat{A}(s)$ that we would have if there were no ratio-estimator bias[4]. Let G denote a population subgroup.

The *random sampling error* is equal to $\sum_{s \in S_E \cap G} \hat{A}(s) - E(\sum_{s \in S_E \cap G} \hat{A}(s))$ and *ratio-estimator bias* is equal to $E(\sum_{s \in S_E \cap G} \hat{A}(s)) - E(\sum_{s \in S_E \cap G} \hat{A}_1(s))$.

The *measurement error bias* or *bias from data error* is defined as $E(\sum_{s \in S_E \cap G} \hat{A}_1(s)) - E(\sum_{s \in S_E \cap G} \hat{A}_1^*(s))$, with $\hat{A}_1^*(s)$ defined as the value of $\hat{A}_1(s)$ if there were no data collection error or data processing error or ratio estimator bias.

Contamination occurs when the A.C.E. selection of a given block cluster alters the way the census is conducted there. Let $A_0(s)$ denote the true adjustment factor if A.C.E. had selected the enumeration's block cluster (or subsampled part of the block cluster). If there is no contamination then $A_0(s)$ will be the same as A(s). For persons whose housing units were sampled in A.C.E., $A(s) = A_0(s)$. The *contamination bias* is defined as $E(\sum_{s \in S_E \cap G} A_0(s)) - E(\sum_{s \in S_E \cap G} A(s))$.

Model bias is defined as $E(\sum_{s \in S_E \cap G} \hat{A}_1^*(s)) - E(\sum_{s \in S_E \cap G} A_0(s))$.

---

[4]Thus, if ratio estimator bias in $\hat{A}(s)$ is b(s), define $\hat{A}_1(s) = \hat{A}(s) - b(s)$.

Observe that the total error in the DSE for a subgroup G of size N_G is

$$\sum_{s \in S_E \cap G} \hat{A}(s) - N_G = \sum_{s \in S_E \cap G} \hat{A}(s) - E(\sum_{s \in S_E \cap G} \hat{A}(s)) \quad \text{random error}$$

$$+ E(\sum_{s \in S_E \cap G} \hat{A}(s)) - E(\sum_{s \in S_E \cap G} \hat{A}_1(s)) \quad \text{ratio-estimator bias}$$

$$+ E(\sum_{s \in S_E \cap G} \hat{A}_1(s)) - E(\sum_{s \in S_E \cap G} \hat{A}_1^*(s)) \quad \text{measurement error bias}$$

$$+ E(\sum_{s \in S_E \cap G} \hat{A}_1^*(s)) - E(\sum_{s \in S_E \cap G} A_0(s)) \quad \text{model bias}$$

$$+ E(\sum_{s \in S_E \cap G} A_0(s)) - E(\sum_{s \in S_E \cap G} A(s)) \quad \text{contamination bias}$$

$$- [N_G - E(\sum_{s \in S_E \cap G} A(s))] \quad \text{number of uncountable people in group G.}$$

It is important to note that the components of error on the right-hand side of the equation define all of the major error components for the DSE. We know that is so because the equation is an identity, and the left-hand side is the total error, the difference between the estimate and the quantity of interest. It is of interest to further break down the major components of error into their constituents. In the next section, we decompose model error into the sum of correlation bias and synthetic estimation error. Part C, following, provides a detailed outline of the components of measurement error.

4. Model Bias

There are two components of model bias: correlation bias and synthetic estimation bias. To distinguish these components, we will introduce some additional notation. Let $p_{U1}(s)$ and $p_{U2}(s)$ be the marginal probabilities of person $s \in S_T$ being missed in the census and P sample, respectively, and let $q_1(s) = 1 - p_{U1}(s)$ and $q_2(s) = 1 - p_{U2}(s)$ be the respective probabilities of being enumerated in the census and the P sample. Define $q_{12}(s)$ as the joint probability of being enumerated in both the census and the P sample. Let k(s) be the countable poststratum of s, i.e., the set of all persons t in the same poststratum as s such that $p_{U1}(t) < 1$. Define means, variances, and covariance for countable poststratum k by $\bar{q}_{kj}$, $\bar{q}_{k12}$, σ^2_{kj}, σ_{k12} respectively, for j = 1, 2,

$$\bar{q}_{kj} = \operatorname*{ave}_{t \in k} q_j(t), \qquad \bar{q}_{k12} = \operatorname*{ave}_{t \in k} q_{k12}(t),$$

$$\sigma^2_{kj} = \operatorname*{ave}_{t \in k} (q_j(t) - \bar{q}_{kj})^2, \qquad \sigma_{k12} = \bar{q}_{k12} - \bar{q}_{k1}\bar{q}_{k2}.$$

The average probability of erroneous enumeration in countable poststratum k is $\bar{p}_{k,EE} = \underset{t\in k}{\text{ave}}\, p_{EE}(t)$.

Further, define the average census-enumeration probability and the average erroneous-enumeration probability within the part of countable poststratum k that is in group G by $\bar{q}_{Gk(s)1}$ and $\bar{p}_{Gk(s)EE}$, respectively. Now define the post-stratum-level factor $a_{k(s)}$ and the group G poststratum-level factor $A_{Gk(s)}$ by $A_{k(s)} = (1 - \bar{p}_{k(s),EE})/\bar{q}_{k(s)1}$ and $A_{Gk(s)} = (1 - \bar{p}_{Gk(s),EE})/\bar{q}_{Gk(s)1}$.

We now make some preliminary observations before decomposing model bias into the sum of correlation bias and synthetic estimation error. First, recall that $\hat{A}_1^*(s)$ is not subject to data bias or ratio estimator bias and notice that $E(\hat{A}_1^*(s)) = (1 - \bar{p}_{k(s),EE})\bar{q}_{k(s)2}/\bar{q}_{k(s)12}$. What we want $\hat{A}_1^*(s)$ to estimate is $A_{k(s)}$, the poststratum-level factor for enumeration s. The expected error is thus $(1 - \bar{p}_{k(s),EE})\bar{q}_{k(s)2}/\bar{q}_{k(s)12} - A_{k(s)} = - A_{k(s)}\psi_{k(s)}$, with

$$\psi_k = \frac{\sigma_{k12}}{\bar{q}_{k1}\bar{q}_{k2} + \sigma_{k12}}.$$

Lemma (Decomposition of Model Bias).

$$\text{model bias} = E\Big(\sum_{s\in S_E\cap G} \hat{A}_1^*(s)\Big) - E\Big(\sum_{s\in S_E\cap G} A_0(s)\Big)$$

$$= E\Big(\sum_{s\in S_E\cap G} \frac{-(1 - \bar{p}_{k(s),EE})\psi_{k(s)}}{\bar{q}_{k(s)1}}\Big) + E\Big(\sum_{s\in S_E\cap G} \frac{1 - \bar{p}_{k(s),EE}}{\bar{q}_{k(s)1}} - \frac{1 - \bar{p}_{Gk(s),EE}}{\bar{q}_{Gk(s)1}}\Big).$$

$$\text{correlation bias} \quad + \quad \text{synthetic estimation bias}$$

Proof. First observe that

$$E\Big(\sum_{s\in S_E\cap G} \frac{1 - p_{EE}(s)}{q_1(s)}\Big) = E\Big(\sum_{s\in S_E\cap G} A_{Gk(s)}\Big).$$

This holds because

$$
\begin{aligned}
E\Big(\sum_{s\in S_E\cap G}\frac{1 - p_{EE}(s)}{q_1(s)}\Big) &= \sum_{s\in S_0\cap G} 1 - p_{EE}(s) \\
&= \sum_k \sum_{s\in S_0\cap G\cap k} 1 - p_{EE}(s) \\
&= \sum_k \sum_{s\in S_0\cap G\cap k} 1 - \bar{p}_{Gk,EE} \\
&= \sum_k \frac{1 - \bar{p}_{Gk,EE}}{\bar{q}_{Gk1}} \sum_{s\in S_0\cap G\cap k} q_1(s) \\
&= \sum_k \sum_{s\in S_0\cap G\cap k} \frac{1 - \bar{p}_{Gk,EE}}{\bar{q}_{Gk1}} q_1(s) \\
&= E\Big(\sum_k \sum_{s\in S_E\cap G\cap k} \frac{1 - \bar{p}_{Gk,EE}}{\bar{q}_{Gk1}}\Big) \\
&= E\Big(\sum_{s\in S_E\cap G} \frac{1 - \bar{p}_{Gk(s),EE}}{\bar{q}_{Gk(s)1}}\Big) \\
&= E\Big(\sum_{s\in S_E\cap G} A_{Gk(s)}\Big).
\end{aligned}
$$

We can now decompose the model bias into the sum of correlation bias and synthetic estimation bias.

$$
\begin{aligned}
\text{model bias} &= E\Big(\sum_{s\in S_E\cap G} \hat{A}_1^*(s)\Big) - E\Big(\sum_{s\in S_E\cap G} A_0(s)\Big) \\
&= E\Big(\sum_{s\in S_E\cap G} \frac{(1-\bar{p}_{k(s),EE})\bar{q}_{k(s)2}}{\bar{q}_{k(s)12}}\Big) - E\Big(\sum_{s\in S_E\cap G} \frac{1-p_{EE}(s)}{q_1(s)}\Big) \\
&= E\Big(\sum_{s\in S_E\cap G} \frac{(1-\bar{p}_{k(s),EE})\bar{q}_{k(s)2}}{\bar{q}_{k(s)12}}\Big) - E\Big(\sum_{s\in S_E\cap G} A_{Gk(s)}\Big) \\
&= E\Big(\sum_{s\in S_E\cap G} \frac{(1-\bar{p}_{k(s),EE})\bar{q}_{k(s)2}}{\bar{q}_{k(s)12}} - A_{k(s)}\Big) \\
&\qquad + E\Big(\sum_{s\in S_E\cap G} A_{k(s)} - A_{Gk(s)}\Big) \\
&= E\Big(\sum_{s\in S_E\cap G} -A_{k(s)}\psi_{k(s)}\Big) \qquad \text{correlation bias} \\
&\quad + E\Big(\sum_{s\in S_E\cap G} A_{k(s)} - A_{Gk(s)}\Big). \qquad \text{synthetic estimation bias}
\end{aligned}
$$

The advantage of the definitions chosen for correlation bias and synthetic estimation bias is that they partition the model error. Alternative definitions could be considered, but the ones used here have some agreeable aspects. (i) If group G is whole population, synthetic estimation bias is zero. (ii) If we consider a set of disjoint groups whose union is the whole population, we see that the sum of synthetic estimation biases for the groups is zero. (iii) If the average census-enumeration probability and average erroneous enumeration probability within a poststratum differs for group G versus the rest of the population, synthetic estimation bias will be present. (iv) If there is a covariance between census-enumeration probabilities and P-sample enumeration probabilities within poststrata, correlation bias will be present. (v) Differential bias for areas with the same observed characteristics arises from synthetic bias rather than correlation bias. Specifically, it follows from the definition of correlation bias that two groups have the same correlation bias if, for every poststratum, the two groups have equal numbers of enumerations.

C. A Closer Look at the DSE

0. Note on Notation and Terminology

Uppercase letter with ˜ indicates quantity subject to sampling error only; uppercase letter with ^ indicates estimated quantity possibly subject to sampling error and nonsampling error. Uppercase letter except N without ˜ or ^ indicates known quantity. Uppercase letter with * or $\tilde{}$* denotes an unobserved target, which is a construct that does not need to be known. A lower-case letter with a ˜ denotes a sampling error and an unadorned lower-case letter denotes a nonsampling error. The subscript P refers to a quantity based on the P sample, E to a quantity based on the E sample, C to a quantity based on the census, and CE (or CP) to a quantity based on the census and the E sample (or P sample). Lower-case symbols without a $\tilde{}$ generally denote parameters for or components of nonsampling error.

Unless explicitly stated, all sample-based quantities other than statistics computed by the Census Bureau for the DSE are weighted by reciprocals of selection probabilities.

1. Description of the Direct DSE and Adjustment Factor

The direct DSE and adjustment factor are calculated *within each poststratum. For simplicity of presentation, poststratum indicators typically are not shown*. DSEs for groups spread across more than one poststratum are linear combinations of the direct DSEs within poststrata. The DSE evaluated here does not include late census returns.[5] It is important to remember that the DSE is describing only the population in housing units, not the institutionalized or group quarters or homeless populations.

The direct DSE is $\hat{N} = (1 - \hat{E}_E/\hat{N}_E)(\hat{N}_C - I_C)\hat{N}_P/\hat{N}_{CP}$. The adjustment factor is $\hat{A} = \hat{N}/\hat{N}_C$.

[5] The total error model looks at average expected error and expected squared error for different areas and subgroups, but typically does not provide precise estimates of error for individual areas. Because the late census returns tend to be concentrated in a small number of areas, and have very minor effect on DSEs for other areas, they are not treated within this formulation of the total error model.

I_C number of whole person imputations in the census, or equivalently, number of people who are not data defined.

$\hat{E}_E$ estimate of weighted number of erroneous enumerations in the census that were included in the E sample; the erroneous enumerations include cases clerically coded or computer coded as having insufficient information for matching (ref: Design Manual section 4.4).

$\tilde{E}_E$ weighted number of census enumerations that were either erroneous or were geocoded outside their E-sample search areas and that were included in the E sample

$\hat{N}_E$ weighted number of E-sample enumerations; the E-sample excludes whole-person imputations

$\hat{N}_C$ census count, not including late returns

$\hat{N}_{CE} = (1 - \hat{E}_E/\hat{N}_E)(\hat{N}_C - I_C)$

$\tilde{N}_{CE} = (1 - \tilde{E}_E/\hat{N}_E)(\hat{N}_C - I_C)$ number of people in the noninstitutionalized, non-homeless population enumerated in the census and geocoded to their E-sample search areas, excluding erroneous enumerations and whole-person imputations[6]

$\hat{N}_{NM}$ weighted estimate of number of nonmovers who were Census-day residents, i.e., persons in P-sample housing units on the day of the P-sample interview who were also there on census day; persons born since census day are excluded

$\tilde{N}_{NM}$ weighted number of Census-day-resident nonmovers who would have been enumerated in the P sample if there were no P-sample fabrication or missing data; note that $\hat{N}_{NM}$ is an estimate of $\tilde{N}_{NM}$

$\hat{N}_{OM}$ weighted estimate of number of outmovers who were Census-day residents, i.e., persons in P-sample housing units who were Census day residents and who moved out of the housing unit between Census day and the day of the P-sample interview

$\tilde{N}_{OM}$ weighted number of Census-day-resident outmovers who would have been enumerated in the P sample if there were no P-sample fabrication or missing data; note that $\hat{N}_{OM}$ is an estimate of $\tilde{N}_{OM}$

$\hat{N}_{IM}$ weighted estimate of number of inmovers, i.e., persons in P-sample housing units who were not resident there on census day but who were born prior to census day and moved into the housing unit by the day of the P-sample interview

---

[6]We are disregarding error arising from imputation or misreporting of demographic characteristics in the census. An additional component could be introduced for such error, but for the present purposes it does not seem necessary.

$\tilde{N}_{IM}$ weighted number of inmovers who would have been enumerated in the P sample if there were no P-sample data collection error or missing data; note that $\hat{N}_{IM}$ is an estimate of $\tilde{N}_{IM}$

$\hat{N}_{P}$ = $\hat{N}_{P,DSE\text{-}C}$ in most poststrata or = $\hat{N}_{P,DSE\text{-}A}$ in poststrata with fewer than 10 sample outmovers

$\hat{N}_{P,DSE\text{-}C} = \hat{N}_{NM} + \hat{N}_{IM}$ estimate of weighted number of people who would have been enumerated in the P sample if there were no P-sample data collection error or missing data; estimate incorporates weighting adjustments for identified whole-household fabrications and whole-household noninterviews and imputations for missing data, and is adjusted for movers

$\hat{N}_{P,DSE\text{-}A} = \hat{N}_{NM} + \hat{N}_{OM}$

$\tilde{N}_{P} = \tilde{N}_{NM} + \tilde{N}_{OM}$

$\hat{M}_{NM}$ weighted estimate of number of P-sample matches for non-movers, i.e., persons in P-sample housing units who were there both on census day and on the day of the P-sample interview

$\tilde{M}_{NM}$ weighted number of P-sample matches that would have been made for non-movers if there had been no P-sample data collection error or missing data or matching error or geocoding error, conditional on the P-sample search areas that were used

$\hat{M}_{OM}$ weighted number of P-sample matches for out-movers; an outmover is a person who was resident in the A.C.E. block on census day but moved prior to the P-sample interview

$\tilde{M}_{OM}$ weighted estimate of number of P-sample matches that would have been made for out-movers if there had been no P-sample data collection error or missing data or matching error or geocoding error, conditional on the P-sample search areas that were used

$\tilde{M}_{IM}$ weighted number of P-sample matches that would have been made for in-movers if matching had been attempted and there were no P-sample data collection error or missing data or matching error or geocoding error, conditional on the P-sample search areas that were used

$\hat{N}_{CP}$ = $\hat{N}_{CP,DSE\text{-}C}$ in most poststrata or = $\hat{N}_{CP,DSE\text{-}A}$ in poststrata with fewer than 10 sample outmovers

$\hat{N}_{CP,DSE\text{-}C} = \hat{M}_{NM} + \frac{\hat{M}_{OM}}{\hat{N}_{OM}}\hat{N}_{IM}$ weighted estimate of number of P-sample matches to the census, adjusted for movers

$\hat{N}_{CP,DSE\text{-}A} = \hat{M}_{NM} + \hat{M}_{OM}$

$\tilde{N}_{CP} \quad = \tilde{M}_{NM} + \tilde{M}_{OM}$ weighted number of P-sample matches to the census that would have been made if matches for outmovers were used and there were no nonsampling measurement errors (P-sample data collection error, missing data, matching error, geocoding error, etc.), conditional on the P-sample search areas that were used

Comment. The estimator $\hat{N}_{CP}$ for the number of matches treats movers inconsistently by mixing the number of inmovers and the match rate for outmovers. The quantity of interest involves outmovers rather than inmovers, as reflected in the definition of $\tilde{N}_{CP}$.

2. Correlation Bias

Correlation bias in the DSE is $\tilde{N}^{*} - \tilde{N}^{**}$.
Correlation bias in the adjustment factor is $a_{\text{correlation bias}} = (\tilde{N}^{*} - \tilde{N}^{**})/\hat{N}_C$.

$\tilde{N}^{*}_{CE}$ expected value of $\tilde{N}_{CE}$

$\tilde{N}^{*}_{P}$ expected value of $\tilde{N}_{P}$

$\tilde{N}^{*}_{CP}$ expected value of $\tilde{N}_{CP}$

$\tilde{N}^{*}$ basic target DSE = $\tilde{N}^{*}_{CE}\tilde{N}^{*}_{P}/\tilde{N}^{*}_{CP}$ what the DSE is aiming at; this is subject to correlation bias

$\tilde{N}^{**}_{C}$ the census enumeration if all blocks were subject to contamination error

$\tilde{N}^{**}$ expected sum of person-level adjustment factors, with summation over all enumerations, if all blocks were subject to contamination bias

N number of people in the population, whether or not included in the census or P-sample population

$\tilde{N}^{*} - \tilde{N}^{**}$ correlation bias

To interpret $\tilde{N}^{**}$, note that if there were no contamination effects, $\tilde{N}^{**}$ would equal the size of the countable population, i.e., the population with non-zero probabilities of being enumerated. It follows from the definition in section B.3 that model bias in the DSE is $\tilde{N}^{*} - \tilde{N}^{**}$. (We are looking at the entire poststratum, so there is no synthetic estimation bias and model bias equals correlation bias.) The model error in the adjustment factor is $(\tilde{N}^{*} - \tilde{N}^{**})/\hat{N}_C$. Although the denominator, $\hat{N}_C$, is random, it is reasonable to consider the error conditional on the census counts, since we are concerned with the particular census that was taken and not with alternative census counts that could have been recorded. From this conditional perspective, the model error in the adjustment factor is a fixed bias.

Demographic analysis provides information about net bias in the DSE, although separating out the correlation bias from the other biases (ratio-estimator bias, measurement error bias, and contamination bias) requires care; see Spencer (2000a). There is no completely satisfactory method of estimating correlation bias at the poststratum level. The models used by Bell (1993) or Haberman, Jiang, and Spencer (1998) may be used to distribute the correlation bias across poststrata, but I do not know how to validate any of the models. Triple systems estimates based on administrative records may be useful for directly estimating correlation bias at the poststratum level.

The formulas for model bias pertain to poststrata or aggregates of poststrata, and not to groups that cut across poststrata. Synthetic estimation bias can be estimated in part or in whole only for a limited set of areas; see Spencer (1999).

3. Contamination Bias

Contamination bias in the DSE is $\tilde{N}^{**} - N$.

Contamination bias in the adjustment factor is $a_{\text{contamination bias}} = (\tilde{N}^{**} - N)/\hat{N}_C$.

The contamination bias in the DSE is $\tilde{N}^{**} - N$ follows from the definition of contamination bias in section B.3. The contamination error in the adjustment factor is $(\tilde{N}^{**} - N)/\hat{N}_C$. Conditional on the denominator, $\hat{N}_C$, the contamination error in the adjustment factor is a fixed bias.

The sum of contamination bias, ratio-estimator bias, measurement error bias, and correlation bias can be estimated at the national level under certain assumptions (Spencer 2000). If one takes the demographic analysis (DA) totals for age, race, and sex as correct, one can use the difference between DA and the DSE to estimate the net bias for the group. Given separate estimates of ratio estimator bias and measurement error bias, one can subtract them from the net bias estimate to obtain an estimate of the sum of the biases from model error and from contamination. The drawback to this approach is that the DA estimates of totals are considered less reliable than DA estimates of sex ratios.

To independently quantify the extent of contamination bias, we can proceed with two strategies. The Bureau's plans (as described in Evaluation Project N1) are to compare A.C.E. and non-A.C.E. blocks on variables. Those analyses may provide some insight. In addition, one may look at the ratios of census counts on A.C.E. blocks to those on non-A.C.E. blocks. The ratios should be considered at the evaluation-poststratum level and, sample size permitting, for poststratum groups.[7] To see why this is useful, recall the notation of section B, where A(s) is the adjustment factor for enumeration s and $A_0(s)$ is the adjustment factor if contamination bias were present, S_E is the set of enumerations, and S_0 is the set of potential enumerations. The contamination bias in the DSE is equal to the expectation of

[7] A poststratum group is obtained by collapsing poststrata by age and sex. If there is evidence of contamination bias at the level of evaluation poststrata or poststratum groups, it should be modeled at the level of poststrata. Contamination bias may be modeled separately or it may be added in to a component of net bias and the modeling may focus on the net biases. The modeling may proceed along the lines of sections 4.6 and 4.7.

$\sum_{S_E} (A_0(s) - A(s))$, which may be written as $R_0 \sum_{S_E} A(s)$, with $R_0 = \sum_{S_E} A_0(s) / \sum_{S_E} A(s) - 1$. To estimate the bias, we focus on the two pieces separately. Starting with the simpler piece, we may estimate $\sum_{S_E} A(s)$ by the DSE.[8] To estimate R_0, note that if the post-stratification model holds – if $A_0(s)$ and A(s) are constant within poststrata – then $R_0 = A_0/A - 1$. Let $\hat{N}_{C,ACE}$ denote the sample-weighted number of census enumerations in the poststratum in the A.C.E.
From Section B.1 we know that $A_0\hat{N}_{C,ACE}$ approximates the true size of the poststratum, as does $A\hat{N}_C$. Thus, we may estimate R_0 by $\hat{N}_C/\hat{N}_{C,ACE} - 1$.

The accuracy of this estimate depends on homogeneity of the adjustment factors within poststrata. This analysis should be done by evaluation poststratum and, sample sizes permitting, by poststratum. In addition to looking at ratios of numbers of enumerated persons per block cluster, it will be useful to look at the ratio of numbers of enumerated persons per housing unit and the ratio of numbers of housing units per block cluster, to see if they are the same for A.C.E. and non-A.C.E. blocks.

4. Measurement Error and Sampling Error

4.1. Overview

Measurement and sampling error in the DSE is $\hat{N} - \tilde{N}^*$. Measurement and sampling error in the adjustment factor is $a_{m\&s\ error} = (\hat{N} - \tilde{N}^*)/\hat{N}_C$ and is equal to $a_{ratio\ bias} + a_{measurement\ bias} + a_{m\&s\ random\ error}$.

The random component of measurement and sampling error is the sum of two pieces, $\hat{N}/\hat{N}_C - E(\hat{N}/\hat{N}_C)$ and $-[\tilde{N}^*/\hat{N}_C - E(\tilde{N}^*/\hat{N}_C)]$. From a conditional viewpoint, the second piece is zero. Thus, we define $a_{m\&s\ random\ error} = \hat{N}/\hat{N}_C - E(\hat{N}/\hat{N}_C)$; this reflects sampling variance and errors for adjustment for missing data, as well as other random nonsampling errors (such as response variance). The systematic component of sampling error in the adjustment factor is the ratio-estimator bias,

$$a_{ratio\ bias} = E(\hat{A}) - \frac{E(\hat{N}_E) - E(\hat{E}_E)}{E(\hat{N}_E)} \cdot \frac{E(\hat{N}_C) - I_C}{E(\hat{N}_C)} \cdot \frac{E(\hat{N}_P)}{E(\hat{M}_{NM}) + E(\hat{N}_{IM})E(\hat{M}_{OM})/E(\hat{N}_{OM})}.$$

(Note: this formulation assumes that there is no ratio-estimator bias in the components, $\hat{N}_E$, $\hat{E}_E$, $\hat{N}_C$,

---

[8]This DSE is obtained from the A.C.E. blocks only. That is, the adjustment factors are multiplied by the sum of the A.C.E. direct sample weights for census enumerations falling within the A.C.E. sample. Also, although we could adjust the DSE for estimated measurement bias, the effect would be relatively minor.

$\hat{N}_P$, $\hat{M}_{NM}$, $\hat{N}_{IM}$, $\hat{M}_{OM}$, $\hat{N}_{OM}$.) Measurement error bias $a_{\text{measurement bias}}$ is defined as

$$a_{\text{measurement bias}} = \frac{E(\hat{N}_E) - E(\hat{E}_E)}{E(\hat{N}_E)} \cdot \frac{E(\hat{N}_C) - I_C}{E(\hat{N}_C)} \cdot \frac{E(\hat{N}_P)}{E(\hat{M}_{NM}) + E(\hat{N}_{IM})E(\hat{M}_{OM})/E(\hat{N}_{OM})} - \frac{E(\tilde{N}^*)}{E(\hat{N}_C)}.$$

There are two aspects of random sampling error for the DSE as a whole, ratio-estimator bias arising from nonlinearity of the DSE, and variance. The former may be estimated by a jackknife bias estimate, and the latter by a jackknife variance estimate, both obtainable from VPLX. The jackknife variance estimate is unconditional on the census counts $\hat{N}_C$, but that is appropriate because the numerator of the adjustment factor is correlated with $\hat{N}_C$.

Below, we decompose the measurement error so that the nonsampling biases may be estimated from the evaluation data. We begin (section 4.2) with a background discussion of the effect of inconsistent assignments to poststrata based on the P-sample and the E-sample. Then (sections 4.3 - 4.5) we discuss the components of measurement error affecting the estimates of the size of the P-sample population, the E-sample population, and the matched population. A separate component of error for balancing error is not needed because the errors from inconsistent P-sample and E-sample search areas are reflected in components of error for field work (in the E sample) and matching error (in the P sample).

4.2. Effects of errors in inconsistency of poststrata

The classification of a person into a poststratum can be different in the census and the P-sample component of the Accuracy and Coverage Evaluation (A.C.E.), as shown by Salganik (1999). This may cause a bias in the dual system estimate (DSE) because the coverage factors for gross undercount[9] are derived for poststrata based on the P sample and are applied to the poststrata based on census enumerations, or what we will call E-sample poststrata.

The adjustment factor for a poststratum is the product of two factors. The first factor is an adjustment for erroneous enumerations and non-data defined persons, $(1 - \hat{E}_E/\hat{N}_E)(\hat{N}_C - I_C)/\hat{N}_C$, and it involves only E-sample poststrata. The second factor is an adjustment for persons who are not enumerated (including not data defined) in the census. This factor, which is estimated by $\hat{N}_P/\hat{N}_{CP}$, involves only P-sample poststrata. Ideally, this factor would be based on E-sample poststratification, but of course that is not completely feasible.

To understand the bias, it is useful to consider that *each* person enumerated in the P sample *could* be enumerated both ways and assigned to a poststratum two ways, based on either the P-sample data or the census data. Some imagination is required for a person enumerated in the P sample but not the census, and we must make some assumptions to estimate what the E-sample poststratum is for such a

[9]Persons not data defined are treated as part of the gross undercount.

person.[10]

We index the E-sample and P-sample poststrata by h and k, and assume that the indexing is consistent, so that if the variables recorded for a person were consistent between the census and the P sample, and the person were in E-sample poststratum h, the person would also be in P-sample poststratum h. Let $f_G(h|k)$ denote the proportion of group G persons enumerated in P-sample poststratum k who belong to E-sample poststratum h. Let G denote a subgroup of P-sample enumerations, such as enumerations classified as inmovers, outmovers, or nonmovers. Define $\hat{N}_{P,G}(h,k)$ as the number of P-sample enumerations from group G that are in E-sample poststratum h and P-sample poststratum k, and define $\hat{N}_{CP,G}(h,k)$ as the number of P-sample matches from group G belonging to E-sample poststratum h and P-sample poststratum k. The Census Bureau's estimate of the P-sample population size for group G in poststratum i is, say, $\hat{N}_{P,G}(i)$, and because it is based on P-sample poststratification it is equal to $\sum_h \hat{N}_{P,G}(h,i)$. If the estimate were based on E-sample poststratification, it would be $\sum_k \hat{N}_{P,G}(i,k)$. The error from the inconsistent poststratification, say n_{Gr}, is thus

$$n_{Gr} = \hat{N}_{P,G}(i) - \sum_k \hat{N}_{P,G}(i,k).$$

If we had available an estimate of f_G, say $\hat{f}_G$, and if there were no other nonsampling errors in $\hat{N}_{P,G}(i)$'s, we could estimate the expected value of n_{Gr} by

$$\hat{N}_{P,G}(i) - \sum_k \hat{f}_G(i|k)\hat{N}_{P,G}(k). \qquad (*)$$

To allow for the presence of other nonsampling errors in $\hat{N}_{P,G}(i)$, say $n_{P,G}(i)$, and to avoid double counting of errors, we estimate n_{Gr} by

$$\hat{N}_{P,G}(i) - n_{P,G}(i) - \sum_k \hat{f}_G(i|k)(\hat{N}_{P,G}(k) - n_{P,G}(k)). \qquad (**)$$

Notice that the difference between (*) and (**) is second order.

Consider now the question of estimation of $f_G(h|k)$. One set of issues relates to data. For P-sample matches, we have observations on the joint P- and E- sample poststratification, but for P-sample nonmatches we do not have data. Thus, we will want to use the data on the matches to make inferences on the joint classification for the nonmatches. This raises two potential concerns.

[10]Specifically, we mean what E-sample poststratum would have been assigned if the person had been enumerated in the census. It is not unreasonable to view the unknown E-sample poststratum for an unenumerated person as a random variable, but in light of the available information it is most reasonable for the present purposes to view it as fixed. The reference in all cases is to Census-day characteristics, and is distinct from the differences between inmover characteristics and outmover characteristics, as the latter differences reflect the effect of change over time.

Concern 1. Inconsistent or incomplete reporting of demographic characteristics could cause false nonmatches, and hence the inconsistency might be higher for nonmatches than for matches. For this to be a significant concern, there would need to be a substantial amount of cases incorrectly classified as nonmatch because the demographic characteristics reported in the P- sample interview are inconsistent with the census reports or because the P-sample reporting is incomplete. If the extent of false nonmatches is sufficiently low, this cannot be a problem.

Concern 2. Persons who are less likely to be enumerated in the census might tend to have their characteristics reported differently between the census and P sample. Even if the differences in reporting by individuals were not systematic, the differences in sizes of the poststrata could cause systematic differences in the estimates of the poststratum sizes.

To accommodate the first concern, we can design the estimation of f_G to allow for differences between cases with complete reporting of P-sample characteristics and those with incomplete data. The second concern can best be addressed through a sensitivity analysis to show how the estimates of bias vary under alternative scenarios.

Another set of issues relates to modeling. For detailed discussion, see Haberman and Spencer (2001).

A related set of issues arises because the set of matches from which $\hat{f}$ is developed includes some out-of-scope cases, such as students in dormitories incorrectly included in the P sample and matched to census enumerations. To the extent that the inconsistency in poststratification is different for such cases than for other cases in the poststratum, $\hat{f}$ will be biased.

4.3. Error in the Estimated Size of the P-sample Population

The estimate $\hat{N}_P$ of the size of P-sample population is affected by whole-household and partial-household fabrications in P sample, imperfect weighting adjustments for whole-household noninterviews and imperfect imputations for missing data, and misclassification of mover status. The use of DSE-C rather than DSE-B as in 1990 complicates the total error analysis because of the effects of misclassification among nonmovers, inmovers, and outmovers. Define

n_{NM}	$= \hat{N}_{NM} - \tilde{N}_{NM}$	nonsampling error in the estimate of number of resident nonmovers
n_{IM}	$= \hat{N}_{IM} - \tilde{N}_{IM}$	nonsampling error in estimate of number of inmovers
n_{OM}	$= \hat{N}_{OM} - \tilde{N}_{OM}$	nonsampling error in estimate of number of outmovers
n_{IO}	$= \tilde{N}_{IM} - \tilde{N}_{OM}$	difference between weighted number of inmovers and resident outmovers in sample blocks
n_P	$= \hat{N}_P - \tilde{N}_P$	nonsampling error in the estimated size of the P-sample population

The formulation of n_P depends on whether DSE-C or DSE-A is being used in the poststratum.

$$n_{P,DSE\text{-}C} = \hat{N}_{P,DSE\text{-}C} - \tilde{N}_P = n_{NM} + n_{IM} + n_{IO}$$

nonsampling error in the estimated size of the P-sample population for poststrata where DSE-C is used

$$n_{P,DSE\text{-}A} = \hat{N}_{P,DSE\text{-}A} - \tilde{N}_P = n_{NM} + n_{OM}$$

nonsampling error in the estimated size of the P-sample population for poststrata where DSE-A is used

We are interested ultimately in n_P. If we could estimate the mean of n_P directly, that would be sufficient. Given the diverse sources of error affecting n_P, it is more convenient to estimate n_{NM}, n_{IM}, and n_{IO} separately.

4.3.1. Nonmovers

The error in the estimated number of nonmovers is affected by component errors from fabrication (n_{NMf}), misclassification of mover status (n_{NMa}), missing data (n_{NMi}), and error from inconsistent poststratification (n_{NMr}).

$$n_{NM} = n_{NMf} + n_{NMa} + n_{NMi} + n_{NMr}$$

As with n_P, it would be sufficient if we could directly estimate the moments of n_{NM}. In practice, it is more convenient to estimate n_{NMf}, n_{NMa}, n_{NMi}, and n_{NMr} separately.

4.3.1.1. Fabrication and nonmovers

n_{NMf} net number of fabricated persons reported as nonmovers; this is the number of fabricated persons reported as nonmovers minus the number of reported nonmovers who were erroneously classified as fictitious in the P sample production.

4.3.1.2. Misclassification of nonmovers

First we describe some more notation. Let the subscripts NM, IM, and OM denote nonmovers, inmovers, and outmovers, respectively. If we prefix the subscript with R or NR it refers to the subset who are Census-day resident or nonresident, respectively; e.g., NRNM refers to nonmovers who are not Census-day residents. If NM or OM is used without a prefix of R or NR, it refers to Census-day residents only, whereas if IM is used without the R or NR prefix, it refers to all inmovers, whether Census-day resident or not. Let

$n_{X,Y}$ weighted number of nonfabricated persons of group Y on sample blocks who are misclassified as in group X, for X and Y taking values RNM, NRNM, NM, RIM, NRIM, IM, ROM, NROM, or OM but $X \neq Y$.

$n_{NMa} = n_{NM,NRNM} + n_{NM,RIM} + n_{NM,NRIM} + n_{NM,ROM} + n_{NM,NROM} - n_{NRNM,NM} - n_{IM,NM} - n_{ROM,NM} - n_{NROM,NM}$
net error in estimated number of resident nonmovers due to misclassification of mover status or

residency status. (If it helps memory, think of the "a" in the subscript as referring to census-day *a*ddress. Note: this error component pertains only to nonfabricated nonmovers.)

4.3.1.3. Missing data and nonmovers

n_{NMi} arises from incorrect weighting adjustments or imputation for missing data in the P-sample questionnaires, as they affect estimated number of nonmovers; correlated with m_{NMi} (see section 4.5.1.1, below)

4.3.1.4. Error from inconsistent poststratification

n_{NMr} error from inconsistent poststratification. Using the notation of section 4.2, we may express this for poststratum i as $\sum_h \hat{N}_{NM}(h,i) - \sum_k \hat{N}_{NM}(i,k)$.

4.3.2. Inmovers

The error in the estimated number of inmovers is affected by component errors from misclassification of mover status and residency status (n_{IMa}), missing data (n_{IMi}), and error from inconsistent poststratification (n_{IMr}).

Observe

$n_{IM} = n_{IMa} + n_{IMi} + n_{IMr}$

$n_{IMa} = n_{IM,NM} + n_{IM,NRNM} + n_{IM,ROM} + n_{IM,NROM} - n_{NM,IM} - n_{NRNM,IM} - n_{ROM,IM} - n_{NROM,IM}$

= net error in estimated number of inmovers due to misclassification of inmovers as nonmovers and outmovers or conversely, whether residents or not (Note: this error component pertains only to nonfabricated inmovers.)

n_{IMi} arises from incorrect weighting adjustments or imputation for missing data in the P-sample questionnaires, as they affect estimated number of inmovers

n_{IMr} error from inconsistent poststratification. Using the notation of section 4.2, we may express this for poststratum i as $\sum_h \hat{N}_{IM}(h,i) - \sum_k \hat{N}_{IM}(i,k)$.

4.3.3. Resident outmovers

The error in the estimated number of resident outmovers is affected by component errors from fabrication (n_{OMf}), misclassification of mover status (n_{OMa}), missing data (n_{OMi}), and error from inconsistent poststratification (n_{OMr}).

$n_{OM} = n_{OMf} + n_{OMa} + n_{OMi} + n_{OMr}$

n_{OMf} net number of fabricated persons reported as outmovers; this is the number of fabricated persons reported as outmovers minus the number of reported outmovers who were erroneously classified as fictitious in the P sample production.

n_{OMa} $= n_{OM,NM} + n_{OM,NRNM} + n_{OM,IM} + n_{OM,NROM} - n_{NM,OM} - n_{NRNM,OM} - n_{IM,OM} - n_{NROM,OM}$

= error in net number of outmovers due to misclassification of nonmovers, inmovers, and nonresidents as resident outmovers or conversely (Note: this error component pertains only to nonfabricated outmovers.)

n_{OMi} arises from incorrect weighting adjustments or imputation for missing data in the P-sample questionnaires, as they affect estimated number of outmovers.

n_{OMr} error from inconsistent poststratification. Using the notation of section 4.2, we may express this for poststratum i as $\sum_h \hat{N}_{OM}(h,i) - \sum_k \hat{N}_{OM}(i,k)$.

4.3.4. Sampling error

$\tilde{n}_P = \tilde{N}_P - \tilde{N}_P^*$ sampling error in $\hat{N}_P$

Comments on Estimation.

n_{NMf}, n_{OMf} The MER and MES together provide estimates of the number of nonmover fabrications. These are weighted numbers, and reflect noninterview weighting adjustments.

n_{NMa}, n_{IMa}, n_{OMa} To estimate moments of n_{NMa}, n_{IMa}, and n_{OMa} it is sufficient to estimate the sum rather than the components. For example, for n_{NMa} we do not require separate estimates of $n_{NM,NRNM}$, $n_{NM,RIM}$, $n_{NM,NRIM}$, $n_{NM,ROM}$, $n_{NM,NROM}$, $n_{NRNM,NM}$, $n_{RIM,NM}$, $n_{NRIM,NM}$, $n_{ROM,NM}$, and $n_{NROM,NM}$, but only of the sums $n_{NM,NRNM} + n_{NM,RIM} + n_{NM,NRIM} + n_{NM,ROM} + n_{NM,NROM}$ and $n_{NRNM,NM} + n_{RIM,NM} + n_{NRIM,NM} + n_{ROM,NM} + n_{NROM,NM}$ or even just the difference between those two sums. The estimates may be based on the best match-code results of the combined MER and MES studies. These results should be based on evaluation-subsample cases (only!) that are resolved and on cases that are unresolved (with respect to residency status) in production. In performing the calculations, here and elsewhere, the production imputation metholodogy needs to be reapplied to the production-unresolved cases in the evaluation subsample. The methodology needs to be applied once, based on production data, and then again, based on evaluation data. To illustrate, consider for example the calculation of $n_{NM,NRNM}$. Based on the evaluation subsample obtain the weighted number of cases that are non-resident nonmovers (NRNM) according to production data (and the production imputation methodology reapplied to the production data for the evaluation subsample). Then find out the weighted number of those cases that are nonmovers (NM) according to the evaluation data (and the production imputation methodology reapplied to the evaluation data for the evaluation subsample). The result is the estimate of the mean of $n_{NM,NRNM}$.

n_{NMi},n_{IMi}, n_{OMi} These will be estimated from sensitivity analyses, as described in Spencer (2000b).

n_{IO} Use MER data to estimate the weighted number of inmovers; this estimate will be based on the weighted number of inmovers who were A.C.E. interview day residents and living in a housing unit on census day. This estimate may be subject to some error due to misclassification of inmovers who were overseas on census day and from imputation of residency status for persons missing data. Use combined data from the MES and MER to determine the weighted number of outmovers. Some imputation will need to be done for persons with unresolved outmover status. Estimate the mean of n_{IO} by the weighted number of inmovers minus the weighted number of outmovers.

n_{NMr}, n_{IMr}, n_{OMr} Although only the sum of n_{NMr} and n_{IMr} is needed for n_P, we will need n_{IMr} separately for the error in the estimated number of matches. Let $\hat{N}_{NM}(k)$ denote the estimated number of non-movers in any P-sample poststratum k. Using the notation of section 4.2, suppose estimates $\hat{f}_{NM}$ have been developed. Then we may estimate the mean of n_{NMr} for poststratum h by

$$[\hat{N}_{NM}(h) - n_{NMf}(h) - n_{NMa}(h) - n_{NMi}(h)] - \sum_k \hat{f}_{NM}(h|k)[\hat{N}_{NM}(k) - n_{NMf}(k) - n_{NMa}(k) - n_{NMi}(k)].$$

Estimates for n_{IMr} and n_{OMr} are derived analogously. The calculations of n_{NMr}, n_{IMr}, and n_{OMr} do not have to be redone for the estimation, based on sensitivity analyses, of imputation error components n_{NMi}, n_{IMi}, and n_{OMi}.

4.4. Error in the estimated number of matchable enumerations

The estimate $\hat{N}_{CE}$ of the number of enumerated people whose enumerations could in theory be matched in the A.C.E. if the A.C.E. were a 100% sample and there were no matching errors is affected by errors in the estimated numbers of erroneous enumerations. The biases in the two alternative estimates will be the same, although the variances might differ.

Define

$n_{CE} = \hat{N}_{CE} - \tilde{N}_{CE}$ nonsampling error in estimate of number of correct enumerations in the census that were included in the E sample

$n_{EE} = \hat{E}_E - \tilde{E}_E$ nonsampling error in estimate of number of erroneous enumerations in the census that were included in the E sample

Observe that

$$n_{CE} = (\hat{N}_C - I_C)(-n_{EE})/\hat{N}_E.$$

The nonsampling error n_{EE} arises during the processing of the E sample if respondents are misclassified as to whether they are correctly or erroneously enumerated in the original enumeration. Some cases are immediately classifiable but others, the “unresolved”, lack sufficient information for deciding whether the case is a match or a nonmatch; a probability of erroneous enumeration or valid enumeration later is imputed to the latter cases. Thus,

$$n_{EE} = c_e + c_i + c_c$$

c_e weighted net number of erroneous enumerations due to processing errors (e.g., matching errors).

c_i weighted number of misclassifications of erroneous enumerations.

c_c weighted net number of misclassifications of erroneous enumerations due to errors in data collection, whether caused by respondent or interviewer

$\tilde{n}_{CE} \quad = \tilde{N}_{CE} - \tilde{N}^*_{CE}$ sampling error in $\hat{N}_{CE}$

Comments on estimation.

c_e , c_c The sum of these two errors can be estimated from the combined MER and MES E-sample components. In determining these two components, use the evaluation subsample and compare the codes based on production data and the production imputation methodology reapplied to the production data for the evaluation subsample with the codes based on evaluation data and the production imputation methodology reapplied to the evaluation data.

c_i This must be estimated as part of the study of imputation and missing-data error, most likely via sensitivity analysis.

4.5. Error in the number of matches between the P sample and the census

The number of matches for a poststratum is based on DSE-C unless the number of sample outmovers is less than 10, in which case it is based on DSE-A.

$\hat{N}_{CP,DSE\text{-}C} = \hat{M}_{NM} + \dfrac{\hat{M}_{OM}}{\hat{N}_{OM}}\hat{N}_{IM}$ weighted number of P-sample matches to the census, adjusted for movers

Under DSE-C, the error in the estimated number of matches is the sum of the error in the estimated number of nonmover matches, a more complicated contribution of error for matches for movers, and discrepancies in the numbers of matches for outmovers and inmovers.

$$n_{CP,DSE\text{-}C} = \hat{N}_{CP,DSE\text{-}C} - \tilde{N}^*_{CP} = (\hat{M}_{NM} - \tilde{M}_{NM}) + (\frac{\hat{M}_{OM}}{\hat{N}_{OM}}\hat{N}_{IM} - \frac{\tilde{M}_{OM}}{\tilde{N}_{OM}}\tilde{N}_{IM}) + \frac{\tilde{M}_{OM}}{\tilde{N}_{OM}}n_{IO}.$$

The estimate under DSE-A is simpler.

$\hat{N}_{CP,DSE\text{-}A} = \hat{M}_{NM} + \hat{M}_{OM}$ weighted number of P-sample matches to the census, adjusted for movers

Under DSE-A, the error in the estimated number of matches is the sum of the error in the estimated number of nonmover matches and the error in the estimated number of outmover matches.

$$n_{CP,DSE\text{-}A} = \hat{N}_{CP,DSE\text{-}A} - \tilde{N}^*_{CP} = (\hat{M}_{NM} - \tilde{M}_{NM}) + (\hat{M}_{OM} - \tilde{M}_{OM}).$$

The error in $\hat{M}_{NM}$ is

$m_{NM} = \hat{M}_{NM} - \tilde{M}_{NM}$, which will be discussed in section 4.5.1, below.

It may provide insight to note that the relative error in $\frac{\hat{M}_{OM}}{\hat{N}_{OM}}\hat{N}_{IM}$ is approximately equal to

$$\frac{\tilde{M}_{OM}}{\tilde{N}_{OM}}\tilde{N}_{IM}\left(\frac{m_{OM}}{\tilde{M}_{OM}} + \frac{n_{IM}}{\tilde{N}_{IM}} - \frac{n_{OM}}{\tilde{N}_{OM}}\right).$$

The error m_{OM} in the number of matches for outmovers is discussed in section 4.5.2. The errors n_{IM} and n_{OM} were discussed in sections 4.3.2 and 4.3.3.

There is no error component for the effect of fabrications, because fabrications in A.C.E. will almost never match E-sample enumerations. Persons not enumerated in the P-sample because fabrications occurred are simply treated as P-sample misses.

4.5.1. Nonmover matches

The error m_{NM} in the number of matches for nonmovers is due to two sources. First, there is error because the numbers of persons subject to matching are incorrect; the relevant numbers are gross numbers and not simply net. Second, error arises in assignment of match status to a case that was attempted, i.e., error in the match rate.

Let the actual match rate for nonmovers *in the evaluation subsample* be denoted by $\hat{\mu}_{NM}$. The actual match rate that should have been calculated (had matches been attempted) for nonmovers in the sample will be denoted by $\tilde{\mu}_{NM}$. Thus, the error in the match rate is $\hat{\mu}_{NM} - \tilde{\mu}_{NM}$. We assume that the probability of error in assignment of match status is constant within a (P-sample) poststratum, even cases subject to inconsistent P-sample and E-sample poststratification. The error in the number of matches for nonmovers can be expressed as the contribution of error in the match rate, error in the number of persons subject to matching, and the product of the two errors (which will be smaller than the other two error contributions):

$$m_{NM} = \hat{\mu}_{NM}\hat{N}_{NM} - \tilde{\mu}_{NM}\tilde{N}_{NM} = (\hat{\mu}_{NM} - \tilde{\mu}_{NM})\hat{N}_{NM} + \hat{\mu}_{NM}n_{NM} - (\hat{\mu}_{NM} - \tilde{\mu}_{NM})n_{NM}.$$

Recall that $\hat{\mu}_{NM}$ and $\hat{N}_{NM}$ are observed in the evaluation subsample, and the estimation of n_{NM} was discussed in Section 4.3.1. What is left is to estimate the error in the match rate, $\hat{\mu}_{NM} - \tilde{\mu}_{NM}$, which will be discussed in Section 4.5.3, below.

4.5.2. Movers

For insight, note that the relative error in $(\hat{M}_{OM}/\hat{N}_{OM})\hat{N}_{IM}$ is approximately equal to

$$\frac{\tilde{M}_{OM}}{\tilde{N}_{OM}}\tilde{N}_{IM}\left(\frac{m_{OM}}{\tilde{M}_{OM}} + \frac{n_{IM}}{\tilde{N}_{IM}} - \frac{n_{OM}}{\tilde{N}_{OM}}\right).$$

The error components n_{IM} and n_{OM} were discussed in sections 4.3.2 and 4.3.3.

$$m_{OM} = \hat{M}_{OM} - \tilde{M}_{OM}$$

As with nonmover matches (section 4.5.1), error in the number of outmover matches arises from error in the numbers of persons subject to matching as well as incorrect assignments of match status.

Let the actual match rate for outmovers *in the evaluation subsample* be denoted by $\hat{\mu}_{OM}$. The actual match rate that should have been calculated (had matches been attempted) for outmovers in the sample will be denoted by $\tilde{\mu}_{OM} = \tilde{M}_{OM}/\tilde{N}_{OM}$. Thus, the error in the match rate is $\hat{\mu}_{OM} - \tilde{\mu}_{OM}$. We assume that the probability of error in assignment of match status is constant within a (P-sample) poststratum, even cases subject to inconsistent P-sample and E-sample poststratification. The error in the number of matches for outmovers can be expressed as the contribution of error in the match rate, error in the number of persons subject to matching, and the product of the two errors (which will be smaller than the other two error contributions):

$$m_{OM} = \hat{\mu}_{OM}\hat{N}_{OM} - \tilde{\mu}_{OM}\tilde{N}_{OM} = (\hat{\mu}_{OM} - \tilde{\mu}_{OM})\hat{N}_{OM} + \tilde{\mu}_{OM}n_{OM}.$$

Note that $\hat{\mu}_{OM}$ and $\hat{N}_{OM}$ are observed from the evaluation subsample and the estimation of n_{OM} was discussed in Section 4.5.2.1. What is left is to estimate the error in the match rate, $\hat{\mu}_{OM} - \tilde{\mu}_{OM}$.

4.5.3. Estimation of error in the match rate

The error in the match rate (either for nonmovers or for movers) is due to five sources: incorrect data for matching, clerical matching error, incorrect imputation for missing data, inconsistent poststratification, and imperfect weighting adjustments for household non-interviews. The first two sources of error apply to cases that were not unresolved in the A.C.E. production matching, whereas the third applies to cases that were observed but unresolved in production matching, and the fourth, inconsistent poststratification, causes an error because the estimated match rate refers to cases in a P-sample poststratum but it should refer to cases in an E-sample poststratum. The fifth source applies to cases that were not observed.

The fifth source of error interacts to a very slight extent with the others, because the non-interview weighting adjustments are applied to both resolved and unresolved cases. If the non-interview rates are small, the interaction can be ignored for practical purposes and fourth source of error may be considered to be independent of incorrect data and clerical matching error.

The effect of data error and clerical matching error can be estimated from the MER and the MES respectively, or their net effect can be estimated from differences between the production matches (for both production-resolved cases studied and for the production-unresolved cases, whose imputations were redone based on production data in the evaluation subsample) and the matches after the MER and MES data were obtained and used jointly (and the imputations were redone, based on the evaluation data). From this, estimates of the true match rate for production-resolved cases are obtained and subtracted from the observed match rate to provide a preliminary estimate of the bias in the match rate. The preliminary estimate of bias does not reflect the effect of inconsistent poststratification. The preliminary estimate of bias for P-sample poststratum h may be expressed as, say, $\hat{\mu}_{NM}(h) - \tilde{\mu}'_{NM}(h)$, for nonmovers. The estimate of matches for nonmovers, adjusted for bias from the first three sources of error, is $\tilde{\mu}'_{NM}(h)\tilde{N}_{NM}(h)$. Adjusting for the fourth source of bias as well leads to an estimate of bias $\hat{\mu}_{NM}(h) - \tilde{\mu}_{NM}(h)$, with

$$\tilde{\mu}_{NM}(h) = \frac{\sum_k \hat{f}_{NM}(h|k)[\tilde{N}_{NM}(k) + n_{NM,r}(k)]\tilde{\mu}'_{NM}(k)}{\sum_k \hat{f}_{NM}(h|k)[\tilde{N}_{NM}(k) + n_{NM,r}(k)]}.$$

Estimates of bias in $\hat{\mu}_{OM}$ are derived analogously.

The effect of imputation error and imperfect non-interview weighting adjustments on the match rate for the production-unresolved-match cases is treated as a bias, but the bias is taken to be random, independent of the other sources of random error, and with mean zero. The variance of the random bias is estimated from a sensitivity analysis. See section 4.9 for further discussion.

4.5.4. Sampling error

$\tilde{n}_{CP} = \tilde{N}_{CP} - \tilde{N}^*_{CP}$ sampling error in $\hat{N}_{CP}$

4.6. Net measurement error bias in the adjustment factor

The estimated adjustment factor is

$$\hat{A} = \frac{\hat{N}_E - \hat{E}_E}{\hat{N}_E}\frac{\hat{N}_C - I_C}{\hat{N}_C}\frac{\hat{N}_P}{\hat{N}_{CP}}.$$

If there were no measurement error, the estimated adjustment factor would be

$$\tilde{A} = \frac{(\hat{N}_E - \tilde{E}_E)}{\hat{N}_E}\frac{(\hat{N}_C - I_C)}{\hat{N}_C}\frac{\tilde{N}_P}{\tilde{N}_{CP}}$$

$$= \frac{(\hat{N}_E - \hat{E}_E + n_{EE})}{\hat{N}_E}\frac{(\hat{N}_C - I_C)}{\hat{N}_C}\frac{(\hat{N}_P - n_P)}{(\hat{N}_{CP} - n_{CP})}.$$

We use estimates of the means of n_{EE}, n_P, and n_{CP}, say $\hat{E}(n_{EE})$, $\hat{E}(n_P)$, and $\hat{E}(n_{CP})$, and estimate the net measurement error bias by

$$\hat{A} - \frac{(\hat{N}_E - \hat{E}_E + \hat{E}(n_{EE}))}{\hat{N}_E} \cdot \frac{(\hat{N}_C - I_C)}{\hat{N}_C} \cdot \frac{(\hat{N}_P - \hat{E}(n_P))}{(\hat{N}_{CP} - \hat{E}(n_{CP}))}. \qquad (*)$$

The direct sample estimates of (*) will have excessively high variance for many (if not all) poststrata. Furthermore, their empirical variance across poststrata will exceed the variance of the expected values, say V. Modeling should be performed to improve the accuracy of the estimates, but unless special steps are taken the empirical cross-poststrata variance of the model-based estimates may be smaller than V. This could lead to understatement of the biases in differences of adjusted estimates for one area versus another. Thus, one goal of the modeling should be to match the cross-poststrata variance of the bias estimates to the estimated variance of the actual biases.

The modeling could be performed separately on each of the components of measurement error bias, i.e., $\hat{E}(n_{EE})$, $\hat{E}(n_P)$, and $\hat{E}(n_{CP})$, or even on their components, e.g., model $\hat{E}(n_{EE})$ by modeling each of its components, c_e, c_i, and c_c separately and then build up to an estimate of the mean of n_{EE}. The drawbacks of such an approach are the sheer amount of modeling involved, as well as the difficulty of matching the cross-poststrata variance of the estimates to the variance of the actual biases. A better alternative is to model the net measurement error bias directly. Optionally, it may be reasonable to model the sum of net measurement error bias and ratio-estimator bias or the sum of net measurement error bias, ratio-estimator bias, and contamination bias.

For estimation of measurement error bias in the DSE, it is probably sufficient to estimate the measurement bias at the level of poststratum groups. A poststratum group is defined as the set of poststrata that differ only with respect to age and sex. For the most part, areas have similar age-sex compositions, and so the bias in the DSE for most areas would vary little whether the poststratum-level biases were constant within the poststratum groups or not. The estimated measurement error biases for age-sex groups may be used for estimating correlation bias, however, and for that reason it may be prudent to allow the estimates of bias to vary by age and sex.

The modeling may proceed along the following general lines.

1. Obtain direct estimates of the measurement error biases at the poststratum level and at the poststratum group level.

2. Estimate their sampling variances and covariances.

3. Estimate the variance of the distribution of the actual measurement error biases, $a_{\text{measurement bias}}$, across poststrata and across poststratum groups.

4. Compare the two variances in step 3 to see whether the actual measurement error biases vary appreciably within poststratum groups.

5. Fit a model for the measurement error biases at the poststratum level or at the poststratum group

level, as appropriate. If the modeling is done for poststratum groups, some of the alternative models that could be considered are the following.

a. Fixed effects for each large poststratum group, and random effects for others

b. Fixed effects for all variables used to define poststratum groups (Race/Hispanic Origin, Owner/Renter, Mail Return Rate, Urbanicity/TEA), including some interaction terms where there is sufficient sample size

If modeling is done for poststrata, the following kinds of models could be considered.

c. Fixed effects for each large poststratum group, random effects for other poststratum groups, and random effects for age and sex (7 categories)

d. Fixed effects for each large poststratum group, random effects for other poststratum groups, and fixed effects for age and sex (7 categories)

e. Fixed effects for all variables used to define poststratum groups, including some interaction terms where there is sufficient sample size, and fixed effects for age and sex (7 categories)

The choice of model will depend in on the sample sizes and the sampling variances of the direct estimates. It is plausible that weighted least squares should be used, with weights inverse to a measure of effective sample size for the poststratum groups or poststrata.

6. Adjust the cross-poststrata or cross-poststratum-group variance of the model-based estimates of measurement error biases to match the estimated variance of the distribution of $a_{\text{measurement bias}}$. The method for doing this has not been determined yet. Louis (1984) and Ghosh (1992) offer methods for constraining the empirical variance of the model-based estimates to match a second moment condition. (The second moment condition would involve matching V if the variances of direct poststratum-level estimates were equal, and it adjusts for unequal variances.) The methods of Louis and Ghosh rest on some assumptions not satisfied in the current context. Specifically, they assume that the direct estimates are conditionally independent (given the parameter values), but in the DSE application the direct estimates for poststrata (and poststratum groups) are correlated due to block-cluster sampling. It is not clear whether the independence assumption is crucial or not.

7. Call the resulting estimates $\hat{a}_{\text{measurement bias}}$.
8. Use jackknife methods to estimate the covariance matrix of the estimates $\hat{a}_{\text{measurement bias}}$ for the different poststrata.

4.7. Ratio-estimator bias

Ratio-estimator bias may be estimated directly at the poststratum level with a jackknife estimate of bias or a bootstrap estimate of bias. The covariance matrix of the estimates of ratio-estimator bias may also be estimated by jackknifing, most likely grouped jackknifing to control computation time. If the ratio-estimator biases are small, they may be added to the direct estimates of measurement error bias and the sum of the two biases may be modeled as in section C.4.6. If the variances of the direct estimates are large, modeling may performed along the following lines.

- Obtain direct estimates of ratio-estimator bias at the poststratum level.
- Estimate their sampling variances and estimate the variance of the distribution of the actual measurement error biases i.e., the estimated variance across poststrata of $a_{\text{ratio bias}}$.
- Fit a model for the ratio-estimator biases at the poststratum or poststratum-group level; the biases will vary inversely with the A.C.E. sample size, and they may possibly vary with other poststratum characteristics.
- Adjust the cross-poststrata variance of the model-based estimates of ratio-estimator bias to match the estimated variance of the distribution of $a_{\text{ratio bias}}$.
- Call the resulting estimates $\hat{a}_{\text{ratio bias}}$.
- Use jackknife methods to estimate the covariance matrix of the estimates $\hat{a}_{\text{ratio bias}}$ for the different poststrata.

4.8. Sampling variance

The covariance matrix of the adjustment factors may be estimated by jackknife methods using VPLX. The sample is large enough so that smoothing of the covariance matrix may be unnecessary, but some review should be performed after the data are in hand. Note that the smoothing would not affect the estimates of undercount or adjustment factors, but could affect inferences concerning risk of adjustment and risk of non-adjustment.

4.9. Missing-data error

Errors from missing data are treated as random effects with mean zero. The variances of the random effects are estimated from sensitivity analyses applied to the cases that are unresolved in the production estimation. The sensitivity analyses being performed will yield alternative vectors of poststratum-level adjustment factors. A covariance matrix for the effect of imputation error on the adjustment factor can be estimated from the sensitivity analyses. This sensitivity analysis will reflect imputation error components n_{NMi}, n_{OMi}, n_{IMi}, and c_i and also the effect of imputation error on the estimated match rates, $\hat{\mu}_{NM}$ and $\hat{\mu}_{OM}$. A covariance matrix derived from a sensitivity analysis of non-interview weighting adjustments will also be derived. The two covariance matrices will be added together, and the sum will be added to the covariance matrix for sampling error to yield a covariance matrix for net random error.

5.0 Net Bias in the DSE and Adjustment Factors

The net bias in the adjustment factor is equal to the sum $a_{\text{correlation bias}} + a_{\text{contamination bias}} + a_{\text{ratio bias}} + a_{\text{measurement bias}}$. Estimation of each of these has been discussed above, and recall from section 2 that at the poststratum level, model bias is the same as correlation bias. Attention should be paid to possible correlations of estimates of bias, both within-poststratum and across poststrata.

The estimates of $a_{\text{measurement bias}}$, which are based on the EFU subsample of the A.C.E., will be slightly correlated with the estimates of $a_{\text{ratio bias}}$ and $a_{\text{contamination bias}}$, which are based on the full A.C.E. sample. To estimate the variance of the sum $\hat{a}_{\text{non-model bias}} = \hat{a}_{\text{measurement bias}} + \hat{a}_{\text{ratio bias}} + \hat{a}_{\text{contamination bias}}$ at the poststratum level, two approaches may be considered. One is to treat $\hat{a}_{\text{measurement bias}}$ as uncorrelated with $\hat{a}_{\text{ratio bias}} + \hat{a}_{\text{contamination bias}}$, which will lead to an underestimate of the variance of the full sum, but not a severe underestimate. The other approach is to estimate the variance of $\hat{a}_{\text{non-model bias}}$ directly. In either case, the covariance matrix for the estimate can be obtained by a (grouped) jackknife procedure or equivalent.

The estimates of model bias are subject to sampling errors and nonsampling errors. If the estimates of correlation bias are obtained without adjustment for other biases (Spencer 2000), the sampling errors might be small. Although the error in estimates of correlation bias at the poststratum level are affected by sampling errors in sex ratios from the A.C.E., those sampling errors can be controlled by reweighting the sample. There are various ways of adjusting the weights, for example, the E-sample totals or sex ratios for various subgroups could be set to match those of the census (Haberman, Jiang, and Spencer 1998). On the other hand, if estimates of correlation bias include adjustments for non-model biases, the uncertainty of the latter could contribute substantially to the variance (Spencer 2000). The adjustments for non-model biases would introduce a correlation between $\hat{a}_{\text{non-model bias}}$ and $\hat{a}_{\text{correlation bias}}$. Depending on the magnitude of the correlation, the sampling variance of the sum $\hat{a}_{\text{non-model bias}} + \hat{a}_{\text{correlation bias}}$ may need to be estimated directly. The nonsampling errors in the estimates of correlation bias will be either judgmental or based on sensitivity analysis. Accordingly, the covariance matrix for the nonsampling errors in the set of poststratum-level estimates of $a_{\text{correlation bias}}$ can be obtained either from judgment or from a sensitivity analysis. The components of nonsampling error in $\hat{a}_{\text{correlation bias}}$ may be taken to be uncorrelated with the other estimates of bias.

6. Synthetic Estimation Error

Synthetic estimation bias can be estimated as described in Spencer (1999), although it is not clear whether the set of areas for which estimates of synthetic estimation bias will be available will be satisfactory.

D. References

Bell, W. R. (1993) Using Information from Demographic Analysis in Post-Enumeration Survey Estimates, *Journal of the American Statistical Association* **88**, 1106-1118.

Ghosh, M. (1992) Constrained Bayes Estimation with Applications, *Journal of the American Statistical Association* **87**, 533-540.

Gilula, Z. and Haberman, S. J. (1994) “Conditional Log-linear Models for Analyzing Categorical Panel Data”, *Journal of the American Statistical Association* **89** 645-656.

Goodman, L. A. (1972) “Some Multiplicative Models for the Analysis of Cross-classified Data,” *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, 649-696. Berkeley: University of California Press.

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 under contract 50-YABC-2-66023 for the Bureau of the Census. Chicago: National Opinion Research Center.

Haberman, S. J., and Spencer, B. D. (2001) "Estimation of Inconsistent Poststratication in the 2000 A. C. E.". Report for the Bureau of the Census. December 17, 2001. Abt Associates: Cambridge, MA.

Hogan, H. (2000) "Accuracy and Coverage Evaluation: Theory and Application", Paper Prepared for the February 2-3, 2000 DSE Workshop of the NAS Panel to Review the 2000 Census. Washington, D.C.: Bureau of the Census.

Louis, T.A. (1984) "Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods", *Journal of the American Statistical Association* **79**, 393-398.

Mulry, M. H. and Spencer, B. D. (1991) "Total Error in PES Estimates of Population: The Dress Rehearsal Census of 1988", *Journal of the American Statistical Association* **86** 839-854, (with discussion 855-863).

Mulry, M. H. and Spencer, B. D. (1993) "Accuracy of the 1990 Census and Undercount Adjustments", *Journal of the American Statistical Association* **88** 1080-1091.

Navarro, A. (2000) "Accuracy and Coverage Evaluation Survey: Targeted Extended Search Plans", DSSD Census 2000 Procedures and Operations Memorandum Series Q-18, Internal Census Bureau Memorandum, January 12, 2000. Washington, D.C.: Bureau of the Census.

Nordmoe, E. (1993) "Entropy-Based Prediction of Categorical Response Variables in Scanner Panel Data". Unpublished doctoral dissertation, Statistics Department, Northwestern University.

Salganik, M. (1999) "Accuracy and Coverage Evaluation Survey: Consistency of Potential Poststratification Variables", DSSD Census 2000 Procedures and Operations Memorandum Series Q-10, Internal Census Bureau Memorandum. September 22, 1999. Washington, D.C.: Bureau of the Census.

Spencer, B. D. (1999) Synthetic Estimation Error. Final Report, Activity 8 under contract 50-YABC-7-66020 for the Bureau of the Census. Cambridge: Abt Associates.

Spencer, B. D. (2000a) Final Report on Correlation Bias Methodology. April 20, 2000, Activity 8 under contract 50-YABC-7-66020 for the Bureau of the Census. Cambridge: Abt Associates.

Spencer, B. D. (2000b) Initial Report on Methodology for Bias and Uncertainty in Missing Data Procedures. May 9, 2000. Activity 8 under contract 50-YABC-7-66020 for the Bureau of the Census. Cambridge: Abt Associates.