

Estimation Methodology and General Results for the Census 2000 A.C.E. Revision II
 Richard Griffin
 U.S. Census Bureau, Washington, DC 20233

1. Introduction¹

The Accuracy and Coverage Evaluation (A.C.E.) Revision II estimates are measures of the net coverage of the Census 2000 household population. The A.C.E. Revision II estimation method relies on Dual System Estimates (DSEs) that incorporate corrections for measurement errors detected by the A.C.E. Revision II measurement error coding operation and the Further Study of Person Duplication. The estimation method is designed to handle overlap of measurement error detected by both studies in order to avoid overcorrecting for measurement error. An additional feature of the estimation method is that the DSEs for adult males are adjusted for estimated correlation bias.

The general form of the standard DSE without correlation bias adjustment is the data-defined census count multiplied by the ratio of the correct enumeration rate to the match rate. The correct enumeration rate is derived from the E-sample, a sample of census enumerations. The match rate is derived from the P-sample, an independent sample which is matched to the census. Both the correct enumeration rate and the match rate are estimated for population groups called post-strata. For A.C.E. Revision II, different sets of post-strata are used to estimate these rates. Double-sampling ratio adjustments from a subsample of the E and P samples referred to as the A.C.E. Revision sample are used to adjust sample estimates to correct for error detected by the A.C.E. Revision II measurement error coding operation.

For the E-sample, correct enumeration rates are estimated for 525 post-strata based on the following characteristics: Proxy Status, Race/Hispanic Origin Domain, Tenure, Household Relationship, Household Size, Type of Census Return (Mailback vs. Non-Mailback), Date of Return (Early vs. Late), Age and Sex. When summarizing across age and sex, there are 93 Full E-sample post-stratum groups.

For the P-sample, match rates are estimated for 480

post-strata based on these characteristics: Race/Hispanic Origin Domain, Tenure, Size of Metropolitan Statistical Area, Type of Census Enumeration Area, Tract Return Rate (Low vs. High), Region, Age, and Sex. When summarizing across age and sex, there are 64 Full P-sample post-stratum groups.

For of the 525 E-sample post-strata, corrections are incorporated into the estimated correct enumeration rate to account for measurement error detected in the measurement coding operation and the Duplicate Study. Likewise, when computing the match rate for each of the 480 P-sample post-strata, appropriate corrections are incorporated into the match rate.

Corrections are made to the adult male post-strata for correlation bias. Correlation bias refers to the tendency for people enumerated in the census to be more likely included in the coverage survey than those missed in the census. Correlation bias results in a downward bias in the coverage estimates. Because the Census 2000 net undercount was, for the first time, close to zero, with a large number of erroneous enumerations, it was crucial to adjust for errors going in both directions (for details see Shores (2003)).

2. Dual System Estimation

The basic form of the DSE without correlation bias correction calculated within a post-stratum is:

$$DSE = (Cen' - II') \frac{CE}{E} \frac{P}{M} \quad (1)$$

where

- Cen' = the census count excluding late adds
- II' = census records with insufficient information for matching (excluding late adds)
- E = E-sample weighted estimate of total persons
- CE = E- sample weighted estimate of census correct enumerations
- P = P-sample weighted estimate of total persons
- M = P-sample weighted estimate of matches to census correct enumerations

The DSE in (1) can be written as a function of the final census count, Cen , which includes Late adds, and the following three rates:

$$DSE = Cen r_{DD} \frac{r_{CE}}{r_M} \quad (2)$$

where

¹This paper reports the results of research and analysis undertaken by Census bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

$$r_{DD} = \frac{(Cen' - II')}{Cen}$$

is the census data-defined rate

$$r_{CE} = \frac{CE}{E}$$

is the E-sample correct enumeration rate

$$r_M = \frac{M}{P}$$

is the P-sample match rate

The interpretation of r_M may be less obvious than the other two; it is the sample weighted proportion of P-sample persons who were also found in the Census. The general independence assumption underlying DSE is that either the census or A.C.E. inclusion probability are the same for all persons in the post-stratum (both are not required). Assuming causal independence, the match rate r_M estimates the probability of census inclusion for the post-stratum.

Equation (2) also gives an interpretation of how the DSE constructs population estimates within a post-stratum.

- Multiply the census count (Cen) by the data defined rate, r_{DD} , to estimate the number of persons who are data -defined and therefore, eligible for inclusion in the E-sample.
- Reduce this product by multiplying it by the estimated probability of correct enumeration for data defined persons, r_{CE} .
- Increase this result by dividing it by the estimated probability of census inclusion, r_M .

The primary tasks in developing DSEs at the post-stratum level are the estimation of the three rates involved. The estimate r_{DD} is straightforward because it's based on 100-percent census tabulations. The development of the estimates r_{CE} and r_M is more challenging and is discussed in detail in Sections 3 and 4, respectively.

The different estimation tasks can be tackled one term at a time. Basically, the goal is to estimate the numerators and denominators of the terms r_{CE} and r_M . Since E, the estimated number of total census enumerations with sufficient information, is a simple, direct sample-weighted estimate, the challenges relate mostly to developing the estimates CE, P, and M. The estimation challenges for A.C.E. Revision II focus on accounting for: (i) information from the revised coding of the A.C.E. Revision Sample, (ii) information from the A.C.E. Revision II study of census duplicates, and (iii) different post-stratification schemes for the Full E- and P-Samples. The most difficult issue is (ii).

Before proceeding to a detailed discussion of the A.C.E. Revision II DSE components, consider the general nature of the estimator. While the basic DSE shown in

equation (1) was applied in the 1990 PES, the March 2001 A.C.E. incorporated the modification called PES-C estimation. This DSE had the general form:

$$DSE^C = (Cen' - II') \frac{CE}{E} \frac{P_{nm} + P_{im}}{M_{nm} + \frac{M_{om}}{P_{om}} P_{im}} \quad (3)$$

where the following are all P-sample weighted estimates for the given post-stratum:

- M_{nm} = estimate of matches to census persons for nonmovers
- M_{om} = estimate of matches to census persons for outmovers
- P_{nm} = estimate of total nonmovers
- P_{om} = estimate of total outmovers
- P_{im} = estimate of total inmovers

Nonmovers, outmovers, and inmovers were defined with reference to their status in the period of time between Census Day (April 1, 2000) and the A.C.E. interview. Nonmovers were those who did not move during this period, outmovers were those persons who moved out of a sample block during this period, and inmovers are those who moved into a sample block during this period. Equation (3) estimated P-Sample matches (M) as the sum of estimated matches among nonmovers (M_{nm}) and estimated matches among movers. The number of mover matches was estimated as the product of an estimated number of movers (P_{im}) and an estimate of the mover match rate (M_{om} / P_{om}). Thus, P-Sample outmovers were used to estimate the mover match rate while P-Sample inmovers were used to estimate the number of movers. This approach implies that $P_{nm} + P_{im}$ should be used for the estimated total of P-Sample persons (P).

Equation (3) can be further expanded to include post-stratification subscripts. The E- and P- Sample post-strata are denoted by subscripts i and j, respectively. The census term was calculated for the cross-classification of i and j post-strata, denoted ij. The DSE formula, using version C for movers, with different post-strata for the E- and P- Samples is:

$$DSE_{ij}^C = Cen_{ij} r_{DD,ij} \frac{\frac{CE_i}{E_i}}{\left[\frac{M_{nm,j} + \frac{M_{om,j}}{P_{om,j}} P_{im,j}}{P_{nm,j} + P_{im,j}} \right]}$$

3. Estimation of r_{CE}

This section discusses the estimation of the correct enumeration rate, $r_{CE} = CE/E$. The E Sample post-strata are denoted by the subscript i. The Revision E

Sample has post-strata denoted by i , where i is based on collapsed post-strata i . This means that the Revision sample post-strata were obtained by collapsing the E-sample post-strata i (Bell 2002). Collapsing was necessary due to sample size limitations of the Revision sample. The correct enumeration rate is written:

$$r_{CE,i} = \frac{CE_i^{ND} f_{1,i'} + \tilde{CE}_i^D}{E_i} \quad (4)$$

The term CE_i^{ND} estimates the number of correct enumerations in the Full E Sample without duplicate links in post-stratum i . This term includes the probability of not being a duplicate, $1 - p_i$.

The component \tilde{CE}_i^D represents the estimated number of correct enumerations in the Full E Sample with duplicate links in post-stratum i which are retained after unduplication. This term includes the probability of being a duplicate, p_i , as well as the conditional probability that an E Sample case is a correct enumeration given that it is a duplicate to another census enumeration outside the A.C.E. search area.

The total weighted number of persons in post-stratum i in the E Sample are denoted by E_i .

The double-sampling ratio factor $f_{1,i'}$ corrects for measurement error based on the Revision E Sample. It is a ratio of an estimate that uses the revised coding (indicated by *) to an estimate that uses the original coding. These adjustments, which are calculated for measurement error post-strata i , are represented by:

$$f_{1,i'} = \frac{CE_{i'}^{ND*}}{CE_{i'}^{ND}}$$

P- and E- Sample cases with duplicate links to census enumerations outside the search area were assigned a nonzero probability of being a duplicate, p_t . P- and E-Sample cases without duplicate links were assigned a p_t of zero (Mule 2002). This probability is usually 0 or 1 for E- and P- Sample cases, but some duplicate links have a value in between, indicating less confidence that the link is representing the same person. These probabilities also apply to the cases in the E- and P-Revision Samples.

Although the duplicate study identified E- and P-Sample cases linking to census enumerations outside the A.C.E. search area, this study could not determine which component of the link was the correct one. Assuming that the linked person does exist, the goal is

to determine which of the two locations is the appropriate place to count the person. Since linked persons may be geographically close or far apart, this has implications for the degree of synthetic error. Thus, it is necessary to specify the following conditional probability:

z_t the probability that an E-sample case is a correct enumeration given that it is a duplicate to another census enumeration outside the A.C.E. search area.

Details of assigning z_t 's are given in Bell (2002).

The components of Equation (4) are defined below:

$$\begin{aligned} \tilde{CE}_i^D &= \sum_{t \in i} W_{\pi,t}^E p_t z_t PR_{ce\pi,t} \\ CE_i^{ND} &= \sum_{t \in i} W_{\pi,t}^E (1-p_t) PR_{ce\pi,t} \end{aligned}$$

$W_{\pi,t}^E$ is the A.C.E. sampling weight for E-sample person t .

$PR_{ce\pi,t}$ is the probability that person t was a correct enumeration in the original production coding. This is either 0 or 1 unless it was not possible to code the E-sample case a correct or erroneous enumeration. In these cases a probability of correct enumeration was imputed.

$$f_{1,i'} = \frac{CE_{i'}^{ND*}}{CE_{i'}^{ND}} = \frac{\sum_{t \in i'} W_{RR,t}^E (1-p_t) PR_{ceR,t}}{\sum_{t \in i'} W_{RR,t}^E (1-p_t) PR_{ce\pi,t}}$$

$W_{RR,t}^E$ is the A.C.E. Revision Sample weight for person t to be used for revision sample coding.

$W_{R\pi,t}^E$ is the A.C.E. Revision II Sample weight for person t to be used with production coding. These two weights can differ slightly due to Targeted Extended Search status and non-interview adjustment.

$PR_{ceR,t}$ is the probability that person t was a correct enumeration in the A.C.E. Revision sample coding.

$E_i = \sum_{t \in i} W_{\pi,t}^E$ is the total weighted number of persons in the E-sample post-stratum i .

4. Estimation of r_m

This section discusses the estimated match rate in Equation (2). While the E-sample post-strata are indexed by i , the P-sample post-strata are indexed by j .

The match rate for post-stratum j is represented as:

$$r_{Mj} = \frac{\hat{M}_j}{\hat{P}_j}, \text{ where } (5)$$

$$\begin{aligned} \hat{M}_j &= M_{nm,j}^{ND} f_{2,j'} + \tilde{M}_{nm,j}^D \\ &+ \left[\frac{M_{om,j} f_{3,j'}}{P_{om,j} f_{4,j'}} \right] (P_{im,j} f_{5,j'} + g(P_{nm,j}^D - \tilde{P}_{nm,j}^D)) \\ \hat{P}_j &= P_{nm,j}^{ND} f_{6,j'} + \tilde{P}_{nm,j}^D + P_{im,j} f_{5,j'} + g(P_{nm,j}^D - \tilde{P}_{nm,j}^D) \end{aligned}$$

On the P-sample side, this study does not identify whether the linked P-sample case is a resident on Census Day. Thus it is necessary to specify h_t , the probability that a P-sample case is a resident on Census Day given that it links to a census enumeration outside the A.C.E. search area.

Details of assigning h_t 's are given in Bell (2002).

The terms in Equation (5) are defined below.

Summation $t \in j$ denotes summation over the A.C.E. P-sample for post-stratum j. Summation $t \in j'$ denotes summation over the Revision Sample for post-stratum j' . Summation is also indicated over nonmovers, outmovers, or inmovers, using either Production (π) or Revision (R) Sample coding.

$$M_{nm,j}^{ND} = \sum_{t \in \text{nonmover} \in j} W_{\pi,t}^P (1-p_t) PR_{res\pi,t} PR_{m\pi,t}$$

is the estimated number of nonmover non-duplicate link persons determined to have been Census Day residents who matched to the census in the P-sample in post-stratum j.

$$\tilde{M}_{nm,j}^D = \sum_{t \in \text{nonmover} \in j} W_{\pi,t}^P p_t h_t PR_{res\pi,t} PR_{m\pi,t}$$

is the estimated number of nonmover duplicate link persons determined to have been Census Day residents who matched to the census.

$$P_{nm,j}^{ND} = \sum_{t \in \text{nonmover} \in j} W_{\pi,t}^P (1-p_t) PR_{res\pi,t}$$

is the estimated number of nonmover non-duplicate link persons determined to have been Census Day residents in the full P-sample in post-stratum j.

$$\tilde{P}_{nm,j}^D = \sum_{t \in \text{nonmover} \in j} W_{\pi,t}^P p_t h_t PR_{res\pi,t}$$

is the estimated number of nonmover duplicate link persons determined to have been Census Day residents.

$W_{\pi,t}^P$ is the P-sample production weight of person t.

p_t is the probability that person t had a duplicate link outside the search area.

$PR_{m\pi,t}$ is the probability that person t was a match in the production coding.

$PR_{res\pi,t}$ is the probability that person t was a resident in the production coding.

$$\begin{aligned} f_{2,j'} &= \frac{M_{nm,j'}^{ND*}}{M_{nm,j'}^{ND}} \\ &= \frac{\sum_{t \in \text{revision nonmover} \in j'} W_{RR,t}^P (1-p_t) PR_{resR,t} PR_{mR,t}}{\sum_{t \in \text{production nonmover} \in j'} W_{R\pi,t}^P (1-p_t) PR_{res\pi,t} PR_{m\pi,t}} \end{aligned}$$

$PR_{mR,t}$ is the probability that person t was a match in the Revision Sample coding.

$PR_{resR,t}$ is the probability that person t was a resident in the Revision Sample coding.

$W_{RR,t}^P$ is the A.C.E. revision sample weight of person t to be used for Revision Sample coding.

$W_{R\pi,t}^P$ is the A.C.E. revision sample weight of person t to be used with production coding. These two weights could differ slightly depending on targeted extended search status and the non-interview adjustment.

$$M_{om,j} = \sum_{t \in \text{outmover} \in j} W_{\pi,t}^P PR_{res\pi,t} PR_{m\pi,t}$$

is the estimated number of matched outmovers in the full E-sample post-stratum j.

$$\begin{aligned} f_{3,j'} &= \frac{M_{om,j'}^*}{M_{om,j'}} \\ &= \frac{\sum_{t \in \text{revision outmover} \in j'} W_{RR,t}^P PR_{resR,t} PR_{mR,t}}{\sum_{t \in \text{production outmover} \in j'} W_{R\pi,t}^P PR_{res\pi,t} PR_{m\pi,t}} \end{aligned}$$

is the double sampling ratio for matched outmovers for post-stratum j' .

$$P_{om,j} = \sum_{t \in \text{outmover} \in j} W_{\pi,t}^P PR_{res\pi,t}$$

is the estimated number of outmovers in the full sample for post-stratum j

$$f_{4,j'} = \frac{P_{om,j'}^*}{P_{om,j'}} = \frac{\sum_{t \in \text{revision outmover} \in j'} W_{RR,t}^P PR_{resR,t}}{\sum_{t \in \text{production outmover} \in j'} W_{R\pi,t}^P PR_{res\pi,t}}$$

is the double sampling ratio for outmovers for post-stratum j' .

$P_{im,j} = \sum_{t \in \text{inmover} \in j} W_{\pi,t}^P$ is the estimated number of in-movers in the full sample for post-stratum j

$$f_{5,j'} = \frac{P_{im,j'}^*}{P_{im,j'}} = \frac{\sum_{t \in \text{revision inmover} \in j'} W_{RR,t}^P PR_{inmoverR,t}}{\sum_{t \in \text{production inmover} \in j'} W_{RR,t}^P}$$

is the double sampling ratio for in-movers for post-stratum j' .

$PR_{inmoverR,t}$ is the probability that person t in the Revision Sample is an in-mover.

$$f_{6,j'} = \frac{P_{nm,j'}^{ND*}}{P_{nm,j'}^{ND}} = \frac{\sum_{t \in \text{revision nonmover} \in j'} W_{RR,t}^P (1-p_t) PR_{resR,t}}{\sum_{t \in \text{production nonmover} \in j'} W_{RR,t}^P (1-p_t) P_{res\pi,t}}$$

duplicate link in post-stratum j' .

$$g(P_{nm,j}^D - \tilde{P}_{nm,j}^D)$$

The term g adjusts the number of in-movers for those full P-sample non-movers who are determined to be non-residents because of duplicate links. Some of these non-residents are non-residents because they are in-movers and should be added to the count of in-movers.

The term $P_{nm,j}^D - \tilde{P}_{nm,j}^D$ is an estimate of non-residents among non-movers with duplicate links. This term is multiplied by g , which is an estimate of the proportion of originally-coded non-movers with duplicate links who are true residents whom have moved since Census day. The term g is estimated using the Revision Sample and both the original A.C.E. and the revision coding as follows:

$$g = \frac{P_{nm,im}^D}{P_{nm,nr}^D}$$

$P_{nm,im}^D$ is an estimate of persons (using the Revision P-sample) with a duplicate link who were originally coded as a non-mover but the revision coding determined them to be in-movers, which are, of course, a subset of non-residents.

$P_{nm,nr}^D$ is an estimate of persons (using the Revision P-sample) with a duplicate link who were originally coded as a non-mover but the revision coding determined them to be non-residents.

A couple of important assumptions are:

- If the revision coding determined a person was a non-resident, they really are a non-resident; i.e., revision coded non-residents are a subset of true non-residents.
- The rate of in-movers for revision-coded non-residents is the same as that for true non-residents.

5. The A.C.E. Revision II DSE Formula

The A.C.E. Revision II DSE formula, using version C for movers, separate E-and P-Sample post-strata, measurement error corrections from the E-and P-Revision Samples and Duplicate Study results is written as:

$$DSE_{ij}^C = Cen_{ij} r_{DD,ij} \frac{r_{CE,i}}{r_{M,j}}, \text{ where } (6)$$

$$r_{CE,i} = \frac{CE_i^{ND} f_{1,i'} + \tilde{CE}_i^D}{E_i}, \text{ and}$$

$$r_{M,j} = \frac{\hat{M}_j}{\hat{P}_j}, \text{ where}$$

$$\hat{M}_j = M_{nm,j}^{ND} f_{2,j'} + \tilde{M}_{nm,j}^D$$

$$+ \left[\frac{M_{om,j} f_{3,j'}}{P_{om,j} f_{4,j'}} \right] (P_{im,j} f_{5,j'} + g(P_{nm,j}^D - \tilde{P}_{nm,j}^D)), \text{ and}$$

$$\hat{P}_j = P_{nm,j}^{ND} f_{6,j'} + \tilde{P}_{nm,j}^D + P_{im,j} f_{5,j'} + g(P_{nm,j}^D - \tilde{P}_{nm,j}^D)$$

PES-C was used because it was easier to count the in-movers than the out-movers in the P-Sample but it was easier to determine the match rate for out-movers. In some small post-strata, the number of in-movers was substantially larger than the number of out-movers. If there were only a few out-movers, the out-mover match rate was subject to high sampling error. In these post-strata it was not considered appropriate to apply a suspect match rate to what could be a relatively large number of in-movers, so PES-A was used. PES-A uses only out-movers. PES-A was applied for post-strata with 9 or fewer P-Sample out-movers. For these post-strata, we made the assumption that some of the duplicate links determined not to have been residents were really out-movers.

The A.C.E. Revision II DSE formula that uses version A for movers with different post-strata for the E- and P-Samples is given in Bell (2002). This formula was used 93 times out of 489 full P-sample post-strata.

The DSE estimates are adjusted to correct for correlation bias. 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 A.C.E. 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 A.C.E. Estimates of correlation bias are calculated using the “two-group model” and sex ratios from Demographic Analysis (DA). The sex ratio is defined as the number of males divided by the number of females. This model assumes no correlation bias for females or for males under 18 years of age; no correlation bias adjustment for Nonblack males aged 18-29; and that Black males have a relative correlation bias that is different than the relative correlation bias for Nonblack males. The correlation bias adjustment is also done by three age categories: 18-29, 30-49, and 50 and over. This model further assumes that relative correlation bias is constant over male post-strata within age groups. The Race/Hispanic Origin Domain variable is used to categorize Black and Nonblack.

The DA totals are adjusted to make them comparable with the A.C.E. universe and with the A.C.E. Race/Hispanic Origin Domains. In general the correlation factor, c_k , is defined for three age groups k within both the Black and Non-Black Domain such that $E[c_k DSE_k^m] = \text{True male population for age group } k$, where DSE_k^m is the sum of DSEs over male-post-strata in age group k . Since the purpose of this adjustment is to reflect persons missed in both the census and the A.C.E., the value of c_k is not allowed to be less than 1. Further details are given in Shores (2003).

6. General Results

The DSE for a major population group is calculated by summing over the appropriate DSE_y . The net undercount rate, r_{UC} for a major population group is defined as follows:

$$r_{UC} = \frac{DSE - Cen}{DSE}$$

Standard errors are in parentheses.

- The national A.C.E. Revision II percent net undercount estimate for Census 2000 is -0.49 (0.20). This reflects a small estimated overcount of the household population by Census 2000.
- The A.C.E. Revision II estimates that it is likely that Census 2000 experienced a small differential coverage of some major population groups, but this differential is dramatically less than that experienced in the past.
- The estimated percent net undercount for the

Non-Hispanic White Domain is -1.13 (0.20), an estimated overcount of this population, while for the Non-Hispanic Black Domain, the estimate is 1.84 (0.43), an undercount of this population. Only the Non-Hispanic White and Non-Hispanic Black Domains have percent net undercount estimates significantly different from zero.

- The estimated percent net undercount for the Non-Hispanic White Domain is lower than the percent net undercount estimate for the Hispanic Domain.
- The estimated percent net undercount for Non-Hispanic Asians is lower than the Non-Hispanic Black Domain estimate.
- There are no other significant differences among the Race/Hispanic Origin Domain estimates due to sampling variation.
- The A.C.E. Revision II estimate of percent net undercount for Owners is -1.25 (0.20), reflecting an estimated overcount of this population, while for the Non-Owners, the percent net undercount estimate is 1.14 (0.36).
- Among Age/Sex groups, A.C.E. Revision II estimated statistically significant overcounts in Census 2000 of children between 10 and 17 years of age, adult females, and males 50 and over.
- A.C.E. Revision II estimated statistically significant undercounts in Census 2000 for males between 18 and 49 years of age.

7. References

Bell, W.R., Griffin, R., Kostanich, D., Schindler, E and Haines, D. (2002), “Technical Documentation, Chapter 6: A.C.E. Revision II Estimation”, DSSD A.C.E. Revision II Memorandum Series #PP-30r.

Mule, T., Fay, R.E., and Fenstermaker, D. (2003), “Overview and Results of Further Study of Person Duplication for the Accuracy and Coverage Evaluation Revision II, Proceedings of the Survey Research Section.

Shores, R., and Sands, R. (2003), “Correlation Bias Estimation in the Revised Accuracy and Coverage Evaluation, Proceedings of the Survey Research Section.