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

May 1, 2003

DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 3

MEMORANDUM FOR Donna Kostanich
Chair, A.C.E. Revision II Planning Group

From: William R. Bell WRB
Senior Mathematical Statistician for Small Area Estimation

Subject: A.C.E. Revision II: Alternative Options for Tabulating Estimates of Census Correct Enumerations Allowing for Duplicate Links

The attached report discusses various assumptions that can be made when tabulating estimates of census correct enumerations (CEs) to account for cases identified as having duplicate links in the census. The existence of duplicates in the 2000 census E-sample that were not detected and coded as erroneous enumerations (EEs) in the Accuracy and Coverage Evaluation (A.C.E.) survey represents measurement error that led to errors in the March 2001 A.C.E. tabulated estimates of CEs. (These are errors in correcting the errors due to duplication in the census.) Bob Fay's report (Fay, Robert E. (2002), "ESCAP II: Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee for A.C.E. Policy II, Report 9, Revised Version, March 27, 2002) discusses how the census duplicates detected through computer matching provide evidence that the March 2001 A.C.E. underestimated EEs due to census duplicates that were not detected by the A.C.E.

The attached report examines errors that result in tabulations of CEs under various options for treating cases with duplicate links. The errors depend not just on the coded status of cases, but on their true status. Since the latter is unknown, we cannot use data to estimate the magnitude of the errors from the alternative options. Instead, the report examines theoretically the contributions to errors in the tabulations of CEs from groups of census cases defined by their coded status, true status, and whether or not they have a duplicate link. Though we cannot identify the specific cases belonging to these groups, we can make assumptions about which of these groups are likely to be larger than others and which are likely to be smaller, and thus, subject to some additional assumptions, make judgments about which approach to tabulating CEs is likely to lead to the lowest error. The analysis here suggests that the approach used by Fay (2002) and the approach used in A.C.E. Revision II are likely to lead to lower error than the other options considered. The approaches of Fay and of A.C.E. Revision II are related and would produce the same aggregate errors.

Another purpose of the report is to clarify how census duplications lead to error in tabulated estimates of CEs and some limitations of what can be done about this. In particular, in the present setting, it is noted that no unbiased estimate of CEs exists, in the usual sample survey sense.

A preliminary version of this report dated May 21, 2002 was distributed internally. The current version updates some discussion to cover the approach used in the A.C.E. Revision II (though no numerical results from A.C.E. Revision II are presented.) The current version also revises the discussion of dealing with duplicates within the search area.

On Alternative Options for Tabulating Estimates of Census Correct Enumerations Allowing for Duplicate Links

William R. Bell
Revised Version, May 1, 2003 (original version May 21, 2002)

1. Introduction

This report examines, theoretically and under simplifying assumptions, errors that would result in tabulations of correct enumerations under alternative treatments of cases in the census identified as having duplicate links. This is considered separately for duplicate links within the household population and duplicate links between the household population and group quarters (GQ) population. The general approach used here breaks the population down into groups according to whether the person's coded status is CE (correct enumeration) or EE (erroneous enumeration), whether their true status is CE or EE, and whether or not they have a duplicate link. We can then derive the contributions of these groups to the error in tabulations of CEs under alternative options for treatment of the cases with duplicate links.

Section 2 discusses the breakdown of the household (E-sample eligible) population into the groups noted. This leads into the discussion of Section 3 on errors in tabulating CEs that arise from duplications within the household population when using four alternative tabulation options. Option 1 treats the coded status as always correct, ignoring the information from the duplicate links. Option 2 corresponds to the approach used by Fay (2002), which treated all duplicate links as ½ CE (and ½ EE) regardless of whether they were coded CE or EE. A.C.E. Revision II used a different but related approach (Option 4) that maintained the same aggregate totals as Option 2 within certain groups of cases (Fenstermaker and Davis 2002). Option 3, an approach that was considered early on by the A.C.E. Revision II Estimation Subgroup, would have treated all coded EEs as full EEs and all coded CEs with duplicate links as ½ CE (and ½ EE). Section 3 discusses these four options, and examines the errors in their tabulations of CEs in regard to the contributions from the population groups noted in Section 2.

Section 4 deals with duplications between the household and group quarters populations, breaking this set of persons into subgroups according to their coded and true statuses. It then considers tabulation options analogous to most of those of Section 3, plus an additional option that assumes all cases duplicate linked between households and GQs are GQ CEs and household EEs. Fay (2002) and the A.C.E. Revision II (see Fenstermaker and Davis 2002) used this option for these cases. Section 4 examines errors in the tabulations of both household CEs and GQ CEs for the four options in terms of the contributions to error from the subgroups of the cases with household–GQ duplicate links.

Finally, Section 5 briefly considers the issue of overlap between computer detected duplicates (Mule 2002) and duplicates that A.C.E. clerically detected "within the search area" and hence coded as EEs due to duplication. Alternative options for handling this overlap are noted, including the option used by Fay (2002) and in A.C.E. Revision II, which was to restrict the computer detected duplicates used in estimation to those found outside the search area.

The overall goals of this note are to further understanding of differences between the various options considered in regard to errors in estimates of CEs, and also to shed light on the limitations of what can be done to account for duplications in this estimation. Since the groups into which the population is broken down depend on the true status of cases, which is unknown, we cannot use data to estimate the sizes of these groups. We can, however, make assumptions about which groups are likely to be large and which are likely to be small, and thereby determine which tabulation options appear likely to lead to the lowest errors. Section 2 discusses such assumptions for the breakdown of the household population, and Section 4 for the breakdown of cases with duplicate links between the household and GQ population. These are, of course, merely assumptions, not statements of known facts, and other assumptions can be advanced. Under the assumptions proposed here, it appears that the tabulation options used by Fay (2002) and in A.C.E. Revision II (Kostanich 2003) are likely to lead to lower errors than the other options considered. The one possible exception involves cases duplicated between households and GQs for which residents are allowed to claim a usual home elsewhere (a residence other than the GQ). However, even for these cases it is unclear that any alternative tabulation option would produce lower errors.

The analysis presented here makes several simplifying assumptions. The most important of these is the assumption that duplicate links are determined accurately – i.e., that all duplicates are identified and all the duplicates identified truly are duplicates. This assumption, of course, is untrue. The relevance of the analysis here depends not on the assumption being exactly true, but more on the identified duplicates providing reasonably reliable indication, at an aggregate level, of errors in the coded status of cases. If, on the other hand, one believed that the inconsistency observed by Fay (2002) between the estimate of duplicates from computer matching results and the March 2001 A.C.E. estimates of EEs was due more to errors in identification of duplicates than to errors in coded status, then one would be inclined to ignore results of the duplicate study and the analysis here would not be relevant. Since Fay (2002) noted that the duplicate estimates of Mule (2001) were conservative in the direction of possibly underestimating the amount of duplication, and A.C.E. Revision II then estimated even more duplicates (Mule 2002), there is substantial evidence that the inconsistency was indeed due to underestimation of EEs by the March 2001 A.C.E. estimates. One other point worth noting is that if estimates of EEs had been approximately consistent with the duplicate analysis, say within post-strata, then for aggregate results it would not matter much whether we used results of the duplicate analysis or not.

Also for simplicity, and to clarify issues in the treatment of duplicates, this report ignores possible corrections for other forms of measurement error in the E-sample data. In practice, Fay (2002) and A.C.E. Revision II did correct for other errors in the E-sample data in their estimates of 2000 census coverage, and these corrections were coordinated with the corrections for estimates of duplicates.

The presentation here refers to E-sample cases and their coded status. The derivations and discussion for the most part apply generally, and so the "E-sample" could be thought of as the original A.C.E. sample, or the revision E-sample of the A.C.E. Revision II, or to the E-sample arising in any other dual system estimation context. Some analogous issues arise for the tabulation of P-sample estimates of census day residents and matches (Kostanich 2003), but this topic is not considered here.

2. Duplications Within the Household Population – Breaking the Population Into Groups

The approach adopted here is to break cases in the population into groups according to (*i*) whether the E-sample codes them as CE or EE, (*ii*) whether their true status is CE or EE, and (*iii*) whether or not the case has a duplicate link. This breaks the population into 11 distinct groups, as shown in Table 1. From this table we can derive the true number of CEs and compare this to what results from tabulations of CEs under various options. We can then see how the different groups contribute (or not) to the errors from the resulting tabulations. This is done in Section 3. First, in this section I examine the groups in Table 1 and consider what assumptions about their relative sizes seem reasonable. Since the groups depend on the true status of cases, which is unknown, we cannot tabulate actual numbers for these groups, and so must implicitly rely on such assumptions. Thus, the analysis here is theoretical and intended to facilitate understanding about the consequences of different choices of options for treating cases with duplicate links.

Some points worth noting about Table 1 are as follows:

- Table 1 can be thought of as representing the situation if the E-sample were not in fact a sample but a complete recanvas of the population. In this case the entries in the Number of Persons column in Table 1 would be simple counts. Alternatively, these entries can be thought of as the expectations of corresponding sample weighted totals that are (with unbiased weighting) equal to these counts. As noted above, however, we cannot do actual tabulations for these groups since they depend on the unknown true enumeration status of cases. From this perspective the entries in the Number of Persons column represent the expected contributions of these groups towards the true totals of CEs and EEs and to the various tabulations of CEs and EEs.

- Each duplicate link represents one person with two records in the census. The groups in Table 1 that involve duplicate links are thus broken into two subgroups (a and b) that reflect the two census records present for each duplicate link. By definition, corresponding a and b subgroups must contain the same number of persons.

- A second reason we cannot, in practice, tabulate the individual groups 5 to 11 is that they depend on the coded status of both census records for persons with duplicate links. It will be exceedingly rare for persons with duplicate links (outside the search area) to have both of their records be in the E-sample. Thus, in practice we will generally know the coded status of just one of the two records. For groups 5 to 11 the E-sample codes are viewed as the codes that would have resulted for each of the two census records for the duplicate link had that census record been in sample. Thus, Group 5 consists of persons duplicated in the census for whom both of their duplicated records would have been coded CE had they both been in the E-sample.

Table 1. Census Household Enumerations (E-sample universe): Groups Defined by E-sample Coded Status (CE or EE), True Status (CE or EE), and Presence or Absence of a Duplicate Link

Duplicate Link?	Number of Times Enumerated	Group	E-sample Coded Status	True Status	Number of Persons
No	Enumerated once	1.	CE	CE	A
		2.	CE	EE	B
		3.	EE	CE	C
		4.	EE	EE	D
Yes	Enumerated twice, one truly correct	5.a	CE	CE	E
		5.b	CE	EE	E
		6.a	CE	CE	F
		6.b	EE	EE	F
		7.a	EE	CE	G
		7.b	CE	EE	G
		8.a	EE	CE	H
		8.b	EE	EE	H
	Enumerated twice, both truly erroneous	9.a	CE	EE	I
		9.b	CE	EE	I
		10.a	CE	EE	J
		10.b	EE	EE	J
		11.a	EE	EE	K
		11.b	EE	EE	K

Notes to Table 1:

1. Entries in the Number of Persons column can be thought of as simple counts that would result if the E-sample were not in fact a sample but were instead a complete recanvas of the population. Alternatively, these entries can be thought of as the expectations of the sample weighted totals that are unbiased estimates of these counts. See text for more discussion.

2. The groups that involve duplicate links (5 through 11) are each broken into two subgroups (a and b) that reflect the two census records present for each duplicate link. By definition, corresponding a and b subgroups must contain the same number of persons. For simplicity, Table 1 ignores possible triplicates and higher order replication of cases in the census.

3. Since at most one member of a duplicate link can be a true CE, none of the groups in Table 1 that involve duplicate links (groups 5 through 11) have true statuses of CE for both of their subgroups.

4. Table 1 carries an implicit assumption that all duplicate links are found and are accurate (actual duplicates).

- Since at most one member of a duplicate link can be a CE, none of the groups in Table 1 that involve duplicate links (groups 5 to 11) have true statuses of CE for both of their subgroups. Thus, in group 5, which involves duplicate links between two cases both of which are coded CE, subgroup 5.a contains the records that were actually CEs, while subgroup 5.b contains the corresponding records that were actually EEs. However, it is possible for both members of a duplicate link to be actual EEs (regardless of their coded status); this is the case for groups 9, 10, and 11.

- The labeling of the subgroups as a or b is arbitrary. For example, for group 5 I could equally well label the cases with true status CE as the b subgroup and those with true status EE as the a subgroup. Conversely, there is no reason to define another distinct group with the same coded statuses as for group 5 but whose a and b subgroups are labeled in this reverse way. Also, for groups 9 and 11 the assignment of cases to the a and b subgroups is arbitrary, since the coded statuses and true statuses are identical for their two subgroups.

- Table 1 could be defined for subsets of the population as long as any duplicate pairs occur solely *within* the subpopulations under consideration, i.e., for any duplicate pair both their census records are classified and tabulated within the same subpopulation. To the extent that some duplicate pairs have their two records falling in different subpopulations this would have some effect on tabulated estimates of CEs. Since people's true demographic characteristics are uniquely defined, duplication across different demographic subpopulations would occur only due to reporting or processing errors, and such errors should be minor for the very approximate analysis considered here. But duplicates can and will occur across census records in different geographic areas, so this assumption will not hold for subpopulations defined by geography. For such cases one member of the duplicate pair of records is effectively out of the universe under consideration (for the given subpopulation). This situation is somewhat similar to that of duplication between the household and group quarters populations as discussed in Section 4.

- For simplicity, Table 1 ignores possible triplicates and higher order replication of cases in the census.

Some comments on the nature of the groups in Table 1 and their expected relative sizes are as follows:

- Groups 1 to 4 are persons without duplicate links. Group 1 represents those persons correctly enumerated once in the census and accurately coded as such. We expect this to be, by far, the largest group. Group 4 represents those persons included only once in the census but included erroneously and accurately coded as such. Erroneous enumerations include both persons who should not have been included in the census at all (babies born after census day, persons who died before census day, living persons who were not U.S. residents on census day) and persons whom the census counted in the wrong place. For the purposes of A.C.E. and A.C.E. Revision II, "counted in the wrong place" means that the person's true residence was "outside the search area," where the search area includes

the block cluster the person was counted in and, in extended search areas, one additional ring of blocks surrounding this block cluster.

- Groups 2 and 3 involve errors in determination of enumeration status for persons not duplicated. For example, if someone moved after census day and was counted (only) at their new residence, and the E-sample coding did not catch this error, they would be in Group 2. Conversely, if someone moved after census day and was counted (only) at their census day address, but E-sample follow-up could not find evidence that they lived at this address and coded them EE, they would be in Group 3. If the follow-up coding is reasonably successful, Groups 2 and 3 should be small relative to Groups 4 and 1, respectively. We shall see in Section 3 that Groups 2 and 3 have partially offsetting effects on error in tabulations of CEs, but their net effect on error is essentially unavoidable.

- Groups 5 to 8 include persons enumerated twice in the census, once correctly. Examples include persons who moved after census day and were enumerated at both their census day address and their new address, persons who moved before census day and were enumerated at both their pre-census day address and their census day address, persons who maintained two residences on census day and were enumerated at both, and college students enumerated both at their parents' address and in an off-campus address near where they attend school. Groups 5 to 8 would be expected to be the largest groups with duplicate links. The largest of these groups may be Group 6 (F distinct persons), where the enumerations were coded correctly, and Group 5 (E distinct persons), where we failed to detect the erroneous enumeration. Group 7 (G distinct persons) may be smaller. It could include movers where the follow-up conducted months after census day found out there was a move but the respondent (possibly a proxy) misreported the date in relation to census day. It could also include persons with two residences for whom follow-up made the wrong determination of which was the census day address. For aggregate tabulations distinguishing Group 7 from Group 6 is unimportant since both code the correct number of CEs and EEs for their cases. Group 8 (H distinct persons) may be the smallest of these four groups since both enumerations are coded as EE. This could occur if someone who moved after census day was correctly coded as EE at their new address but was also miscoded as EE at their correct census day address.

- Cases in groups 9 to 11 (with numbers of distinct persons I, J, and K) refer to persons included twice in the census but included erroneously both times. One example would be babies born after census day whose family moved after census day but were enumerated, including the new baby, at both their census day address and new address. Another example would be non-U.S. residents visiting the U.S. around the time of the census at more than one place who got enumerated in two different households. We hope these groups of twice erroneously enumerated persons would be small. In addition, for group 9 the E-sample coding fails to detect either of the duplicate enumerations as erroneous, and for group 10 the E-sample coding fails to detect one of the enumerations as erroneous. We thus might hope that group 9 would be smaller than group 10 which would be smaller than group 11, though this is hard to say.

3. Duplications Within the Household Population – Alternative Tabulations of CEs

Keep in mind that we cannot estimate the distinct groups in Table 1, because these depend on the true status of cases, which is unknown. We also cannot use information about whether the duplicate link to an E-sample case is to a census record coded CE or EE, since the linked record is highly unlikely to also be in the E-sample. Thus, we cannot do separate tabulations for coded CEs duplicate linked to coded EEs versus coded CEs duplicate linked to other coded CEs. In fact, I shall assume that the only information we can use in tabulations are the E-sample code of CE or EE, and the presence or absence of a duplicate link. This implies that, for tabulation purposes, essentially all we see is Table 2 below.

Table 2. Census Household Enumerations (E-sample universe): Aggregation of Groups from Table 1 Ignoring True Status, i.e., Defined Only by E-sample Coded Status and Presence or Absence of a Duplicate Link

Contribution from Groups	E-sample Coded Status	Duplicate Link?	Number of Persons
1 and 2	CE	No	A + B
3 and 4	EE	No	C + D
5a,b; 6.a; 7.b; 9a,b; 10.a	CE	Yes	2E + F + G + 2I + J
6.b; 7.a; 8a,b; 10.b; 11a,b	EE	Yes	F + G + 2H + J + 2K

Thus, about all we can do to estimate total CEs is to take a linear combination of the four entries in the Number of Persons column of Table 2. Let the weights assigned to the four entries be w_1, w_2, w_3, and w_4. For the cases without duplicate links it seems hard to argue against assigning weights $w_1 = 1$ to the coded CEs (contribution from Groups 1 and 2) and $w_2 = 0$ to the coded EEs (contribution from Groups 3 and 4.) This leaves us to choose the weights w_3 and w_4. The options discussed below differ in how they assign these weights.

To retract from the preceding point a bit, Michael Beaghen has pointed out that one could use information from the more detailed codes assigned by the E-sample (that indicate the reasons cases were coded as EEs) to break the groups in Tables 1 and 2 further down into subgroups. While this could result in a large number of groups, and thus some complexity, it could be useful *if* it seemed better to treat cases (with duplicate links) differently depending on their detailed codes.

We see from Table 1 that the **true number of correct enumerations** would be

$$CE_{true} = A + C + E + F + G + H .$$

I now consider four options for tabulating estimates of CEs and the errors in these estimates. All use a weight of $w_1 = 1$ for coded CEs without duplicate links (first row of Table 2) and $w_2 = 0$ for coded EEs without duplicate links (second row of Table 2).

Option 1: Treat the E-sample codes as correct and thus ignore the information from the duplicate links. In regard to Table 2 this option uses a weight of $w_3 = 1$ for the third row and $w_4 = 0$ for the fourth row. The tabulation of CEs under this option is

$$CE_1 = A + B + 2E + F + G + 2I + J.$$

The error in this estimate is

$$CE_1 - CE_{true} = (B - C) + E - H + 2I + J. \quad (1)$$

Option 2 (Fay 2002): Treat all cases with duplicate links, whether coded CE or EE, as ½ CE (and ½ EE). This option uses weights of $w_3 = w_4 = ½$ for the third and fourth rows of Table 2. The tabulation of CEs under this option is

$$CE_2 = A + B + ½ [2E + 2F + 2G + 2H + 2I + 2J + 2K]$$

$$= A + B + E + F + G + H + I + J + K.$$

The error in this estimate is

$$CE_2 - CE_{true} = (B - C) + (I + J + K). \quad (2)$$

Option 3: Treat coded CEs with duplicate links as ½ CE (and ½ EE) and treat all coded EEs (with or without duplicate links) as full EEs. The assigned weights are $w_3 = ½$ and $w_4 = 0$ for the third and fourth rows of Table 2. The tabulation of CEs under this option is

$$CE_3 = A + B + ½ [2E + F + G + 2I + J]$$

$$= A + B + E + ½ [F + G + J] + I .$$

The error in this estimate is

$$CE_3 - CE_{true} = (B - C) - ½ (F + G - J) - H + I . \quad (3)$$

Option 4 (A.C.E. Revision II)[1]: Treat coded EEs (with or without duplicate links) as full EEs, i.e., set $w_4 = 0$. Set w_3, the weight on coded CEs with duplicate links, to produce the same aggregate estimate of CEs as for Option 2. Thus,

$$CE_4 = A + B + w_3 [2E + F + G + 2I + J]$$

and $CE_4 = CE_2$ implies that

[1]In A.C.E. Revision II, Option 4 was implemented within population subgroups defined by three Race/Hispanic Origin groups (Blacks, Hispanics, and all others), tenure (renter versus owner), and three linked situations. Also, duplicated persons 18 and older listed as a child of the reference person in just one of the two linked records were handled separately, with the "child of" record considered EE and the "not a child of" record considered CE. See Fenstermaker and Davis (2002) for details.

$$w_3 = [CE_2 \; ! A \; ! B] / [2E + F + G + 2I + J]$$

$$= [E + F + G + H + I + J + K] / [2E + F + G + 2I + J]$$

$$= \tfrac{1}{2} [\text{\# cases with duplicate links}] / [\text{\# coded CEs with duplicate links}].$$

Therefore,

$$CE_4 \; ! \; CE_{true} = CE_2 \; ! \; CE_{true} = (B \; ! \; C) + (I + J + K).$$

The term (B ! C), which appears in the errors of all four estimates, comes from miscoding by the E-sample of cases without duplicate links. It is a nonsampling error unaffected by the treatment of duplicates. Note that these two errors partially offset one another. Note also that the group of persons with true status CE is much larger than the group of persons with true status EE, so we would expect C > B (and thus B ! C < 0), unless misclassification rates for EEs are much higher than for CEs. Contrary to this, however, Martin's (2001) analysis of the evaluation follow-up (EFU) questionnaire identified a specific tendency towards miscoding CEs as EEs. (This was due to coding persons as EE despite failing to record their address when response to EFU was that a person lived elsewhere on census day, since "elsewhere" could have been in the search area, which would imply the case should be coded a CE.) This may suggest that EFU misclassification rates for CEs may actually exceed those for EEs, though this is not a definitive conclusion, since Martin notes EFU may have had other unknown biases, and also revisions to the EFU coding may affect results.

Apart from the term B ! C present in all of equations (1) – (3), we can consider the contributions to error from the other groups for the four estimators.

- The error term for Options 2 and 4 depends only on I, J, and K, which are the contributions from groups 9, 10, and 11. I noted above that we might assume these groups to be smaller than the other groups because they refer to persons included twice in the census and included erroneously both times. If B ! C were small (another assumption) we could then argue that the error under Options 2 and 4 would be small.

- In contrast, the error under Option 1 depends on a contribution of E persons from Group 5, and the error under Option 3 depends on a contribution of ½F persons from Group 6. I noted above that Groups 5 and 6 could be expected to be relatively large. So we have some reason to expect larger errors in tabulations of CEs under Options 1 and 3 than under Options 2 and 4.

It is worth considering the error more generally in terms of any tabulation of CEs with general weights w_1, w_2, w_3, and w_4 for the four rows of Table 2. The general expression for the error in any such tabulation is

$$\{w_1 (A + B) + w_2 (C + D) + w_3 (2E + F + G + 2I + J) + w_4 (F + G + 2H + J + 2K)\}$$

$$- \{A + C + E + F + G + H\}$$

$$= (w_1 - 1)A + w_1 B + (w_2 - 1)C + w_2 D + (2w_3 - 1)E + (w_3 + w_4 - 1)(F + G)$$

$$+ (2w_4 - 1)H + 2w_3 I + (w_3 + w_4)J + 2w_4 K . \qquad (4)$$

Examining equation (4) we can see various conditions required to eliminate from the error the individual terms in (4). Starting with the groups likely to be largest, we can eliminate the terms coming from the various groups as follows:

Group 1, A persons:	$w_1 = 1$	
Group 4, D persons:	$w_2 = 0$	
Group 5, E persons:	$w_3 = ½$	
Group 6, F persons:	$w_3 + w_4 = 1$	$w_4 = ½$
Group 7, G persons:	$w_3 + w_4 = 1$	$w_4 = ½$
Group 8, H persons:	$w_4 = ½$	

We see that we can eliminate A, D, E, and F from the error if we use the weights $(w_1,w_2,w_3,w_4) = (1,0,½,½)$, which is Option 2, the approach used by Fay (2000). These weights also eliminate G and H from the error. Since these terms correspond to the groups of Table 1 that are expected to be relatively large, eliminating these terms from the overall error is important. Use of any other weights allows contributions to error from one or more of the relatively large groups. To avoid a potentially large overall error, any contributions to error from the large groups need to be offset by compensating contributions to error from the smaller groups. This is accomplished by Option 4, since it yields the same tabulated number of CEs as Option 2.

We can draw the following **conclusions for the household population (E-sample eligibles**):

- None of the options for tabulating CEs gives the correct result, CE_{true}. In fact, the general error expression given by equation (4) shows that it is impossible to tabulate the correct number of CEs based on the information actually available (meaning that we have to ignore the true status column of Table 1 and work only with Table 2). In sampling terms, it is impossible to produce an unbiased estimate of CE_{true}.

- If we believe that groups 5, 6, and 7 (E, F, and G persons) are larger than groups 9, 10, and 11 (I, J, and K persons), the latter being the groups of duplicated true EEs, then we expect the estimate of CEs from Option 2 (Fay 2002), which is the same as that from Option 4 (A.C.E. Revision II), to have smaller error than other estimates of CEs, such as those that arise from Options 1 or 3.

4. Allowing for Duplicate Links to Persons in Group Quarters (GQ)

Cases with duplicate links between a person enumeration in a household and a person enumeration in a GQ can be grouped according to the E-sample code for the household enumeration (CE or EE), the true status of the household enumeration (CE or EE), and the true status of the GQ enumeration (CE or EE). Table 3 shows these groups. Since there was no follow-up review of GQ enumerations, Table 3 implicitly assumes all GQ enumerations are effectively coded as CE. The only basis for tabulating any of these groups as something other than CEs within the GQ population would be the information that they duplicate link to household enumerations, the E-sample code for the household enumeration, and available information from the GQ enumerations themselves (such as whether the resident could claim a usual home elsewhere, UHE).

From Table 3 we see that the true number of correct enumerations from these cases within the household population, within the GQ population, and in total would be

$$CE_{HH,true} = L + P \qquad CE_{GQ,true} = M + Q \qquad CE_{TOT,true} = L + M + P + Q.$$

Total (HH plus GQ) CEs is not something used directly in A.C.E. estimation, which inflates estimates of household CEs (by post-strata) to correct for estimated census omissions and then adds in the census GQ results. However, reasoning that cases duplicated between the HH and GQ populations should be counted in only one of these populations not both (except for those that are erroneous both times, which should not be counted in either one), the errors in estimating $CE_{TOT,true}$ give a crude indication of the aggregate effect of the errors arising from these duplicates.

Table 3. Census Household Enumerations with Duplicate Links to GQ Enumerations: Groups Defined by E-Sample Code for Household Enumeration (CE or EE), True Status of Household Enumeration (CE or EE), and True Status of GQ Enumeration (CE or EE)

Group	Household Enumeration Coded Status	Household Enumeration True Status	GQ Enumeration True Status	Number of Persons
GQ-1	CE	CE	EE	L
GQ-2	CE	EE	CE	M
GQ-3	CE	EE	EE	N
GQ-4	EE	CE	EE	P
GQ-5	EE	EE	CE	Q
GQ-6	EE	EE	EE	R

Notes to Table 3 (many are analogous to certain notes to Table 1):

1. Entries in the Number of Persons column can be thought of as simple counts that would result if the E-sample were not in fact a sample but were instead a complete recanvas of the population. Alternatively, these entries can be thought of as the expectations of the sample weighted totals that are unbiased estimates of these counts. However, as in Table 1, we cannot do actual tabulations for these groups since they depend on the true unknown enumeration status of cases.

2. For simplicity Table 3 ignores possible triplicates and higher order replication of cases anywhere between the census household and GQ enumerations (e.g., someone enumerated in two different households and one GQ.)

3. Table 3 could be defined for subpopulations such as post-strata, as long as any duplicate pairs occur solely *within* the subpopulations under consideration. It could also be defined separately for GQs that do not allow reporting of a UHE versus those that do allow reporting of a UHE.

4. Table 3 is assumed to cover all duplicate links between household records and GQs, including those within the search area. Since E-sample follow-up did not search GQ records for duplicates of E-sample enumerations (Childers 2000, pp. 40-47), those E-sample cases that did duplicate GQ enumerations would not have been coded EE for reason of being a duplicate (unless they also duplicated a household enumeration in the search area). They may have been coded as EE for another reason, however, such as having their residence elsewhere (outside the search area) on census day.

5. Since at most one member of a duplicate link can be a CE, none of the groups in Table 3 have true statuses of CE for both the household and GQ enumerations.

6. Table 3 carries an implicit assumption that all duplicate links between household and GQ enumerations are found and are accurate (actual duplicates).

We now examine the contributions to errors from these groups to tabulations of CEs for the household and GQ populations under various options. The tabulations for the total (household plus GQ) population, and hence the corresponding errors, are the same under all the options, and thus are shown only for Option 1. Options 1–3 are analogous to the same options considered before to account for duplicate links within the household population. Option 4 was used by Fay (2002) and A.C.E. Revision II in dealing with duplicate links to GQs.

Option 1: Treat all E-sample codes as correct: cases coded household CEs are tabulated as household CEs and GQ EEs, and cases coded household EEs are tabulated as household EEs and GQ CEs.

$$CE_{HH,1} = L + M + N \qquad CE_{HH,1} - CE_{HH,true} = M + N - P$$

$$CE_{GQ,1} = P + Q + R \qquad CE_{GQ,1} - CE_{GQ,true} = P + R - M$$

$$CE_{TOT,1} = L + M + N + P + Q + R \qquad CE_{TOT,1} - CE_{TOT,true} = N + R$$

Option 2: Treat all enumerations in Table 3 as ½ CE and ½ EE for both the household and GQ populations.

$$CE_{HH,2} = \tfrac{1}{2}(L + M + N + P + Q + R) \qquad CE_{HH,2} - CE_{HH,true} = \tfrac{1}{2}[(M + N + Q + R) - (L + P)]$$

$$CE_{GQ,2} = \tfrac{1}{2}(L + M + N + P + Q + R) \qquad CE_{GQ,2} - CE_{GQ,true} = \tfrac{1}{2}[(L + N + P + R) - (M + Q)]$$

Option 3: Treat all household coded CEs in Table 3 as ½ household CE and ½ GQ CE, and treat all household coded EEs as full household EEs (and thus full GQ CEs).

$$CE_{HH,3} = \tfrac{1}{2}(L + M + N) \qquad CE_{HH,3} - CE_{HH,true} = \tfrac{1}{2}(M + N - L) - P$$

$$CE_{GQ,3} = \tfrac{1}{2}(L + M + N) + P + Q + R \qquad CE_{GQ,3} - CE_{GQ,true} = \tfrac{1}{2}(L + N - M) + P + R$$

Option 4 (Fay 2002 and A.C.E. Revision II): Treat all enumerations in Table 3 as GQ CEs and household EEs.

$$CE_{HH,4} = 0 \qquad CE_{HH,4} - CE_{HH,true} = -(L + P)$$

$$CE_{GQ,4} = L + M + N + P + Q + R \qquad CE_{GQ,4} - CE_{GQ,true} = L + N + P + R$$

Since all four options yield the same value of CE_{TOT}, the choice of option does not affect the total census tabulation of these cases. However, since the choice of option does affect how this population is allocated between the household and GQ populations, and since only the household population was adjusted for coverage errors (census omissions and EEs) in the original A.C.E. and in the A.C.E. Revision II estimation, the choice of option does matter for census coverage estimation.

By making some assumptions about the relative size of the Number of Persons entries in Table 3 we can draw two potential conclusions from the error expressions:

- If the GQ enumeration produces relatively few erroneous enumerations, then in Table 3 groups GQ-2 and GQ-5 would be expected to be the largest, with the corresponding numbers of persons M and Q expected to be larger than L, N, P, and R. In this case Option 4 seems attractive because it is the only option whose error terms do not depend on either M or Q. This seems especially relevant for cases duplicated in GQs that do not allow claiming a UHE.

- For cases duplicated in GQs that do allow claiming of a UHE there may be more question as to how many of those cases should in truth be regarded as resident on census day in a housing unit rather than in a GQ. For these cases it could be that groups GQ-1 and GQ-4 are the largest, so that L and P are larger than M, N, Q, and R. Some accuracy in coding the household record of such cases would imply L > P. Under these conditions the size of the errors for the tabulations options could run in their numerical order, leading us to prefer Option 1 over Option 2 over Option 3 over Option 4. Fay (2002) notes, however, that just because persons in some GQs can claim UHE does not imply that it would be correct for them to do so, that is, not all such persons would actually be resident in the household rather than the GQ population.

To some extent this section, and more generally the coverage estimation of A.C.E. and A.C.E. Revision II, used an assumption that GQ enumeration produces relatively few erroneous enumerations. Rick Griffin has pointed out that this may not be a good assumption. The implications for tabulations of CEs of having possibly significant numbers of EEs in GQs deserves further attention.

5. Handling Persons With Duplicate Links Within the Search Area

The processing and follow-up conducted for the A.C.E. could not identify all E-sample persons duplicated in the census since the follow-up covered only the A.C.E. search areas for the sample block clusters. In most cases the search area was the sample block cluster, but it also included one ring of surrounding blocks in clusters chosen for extended search (Jones 2003). In

either case, E-sample persons with duplicates outside their respective search areas could only be identified as CE or EE according to whether the information collected suggested that they were actually census day residents or not at their E-sample addresses.[2]

On the other hand, within the search area census duplicates of E-sample cases could be identified as duplicates by A.C.E., and an evaluation suggested that A.C.E. did a good job at identifying those duplicates that it searched for (Bean 2001, p. iii). (It searched for duplicates within sample block clusters, but not among all cases in surrounding blocks included in extended search areas. See the note at the end of this section.) Fay (2001) and Mule (2002) in fact used the A.C.E. clerical results on duplication within the search area as a standard to evaluate the efficiency of computer detection of duplicates. Questions thus arise in regard to how to best combine results from computer detection of duplicates with results from the A.C.E. or A.C.E. Revision II coding for those duplicates found within the search area. This is a substantial group – about 2 million persons in Census 2000 (Feldpausch 2001, p. 4) – so this issue is worthy of attention.

We can view the group of persons who are EEs because they are duplicated within the search area in one of three ways:

a. We can regard them as coded EEs but not duplicate links, hence defining "duplicate link" in Table 1 to mean a duplicate outside the search area. Fay (2002) and A.C.E. Revision II (Kostanich 2003) took this approach.

b. We can regard them as coded EEs with duplicate links.

c. We can regard them as duplicate links but not coded EEs, hence defining "coded EE" in Table 1 to mean erroneous for a reason other than being a duplicate within the search area.

The first and third approaches address overlap between coding and computer detection of duplicates within the search area by redefining "computer detected duplicates" or "coded EEs," respectively, to eliminate the overlap. In contrast, the second approach would require that the method used for tabulating CEs address the overlap. Also, an issue that arises with the third approach for the 2000 A.C.E. data is that the information available from follow-up does not permit determination in all cases of whether a case coded as a duplicate within the search area would otherwise have been coded as a CE or EE (Childers 2000 and personal communication). Thus, Option 4 of Section 3 for tabulating CEs could not be applied to these particular A.C.E. cases with duplicate links.

---

[2]Jones (2003, p. 5) notes that, "Person followup found that some E-sample and P-sample persons were Census Day residents of an address outside of the search area. The E-sample person was coded as an 'other residence' erroneous enumeration instead of duplicate and the P-sample person was removed."

Note that the true "enumeration status" of records with duplicate links within the search area is somewhat ambiguous. Although only one of the two linked records can be the actual census day address of the person, since the other record is within the search area it could still be considered a correct enumeration in terms of the definition used for dual system estimation. (This is except for persons that are erroneous for some other reason, such as babies born after census day who were included twice in the census in the same block.) Since the person can only be counted once, the true enumeration status of the pair could be regarded as (CE,EE), but either of the linked records could be considered the correct one. Alternatively, we could consider both of the records to be ½ CE and ½ EE. Assuming random assignment of correct status in the first instance (with probability ½ of assigning the CE code for each of the two census records) these assignments would have, in expectation, the same effect on tabulations of CEs.

The analogous issue with respect to the coding of census cases found to be duplicated within the search area was addressed via coding rules that assigned the CE code to one record from the duplicate pair and the EE code to the other (Childers 2001, pp. 39- 40). These duplicate pairs can thus be regarded as contributing on average one CE and one EE. The exception, as above, would be cases coded as erroneous for some reason other than duplication, in which case the pair would be coded as (EE,EE). If we assume that duplicate links within the search area were accurately determined by A.C.E. (see the Note below for a qualification), then a pair of duplicate records within the search area would never be coded as (CE,CE).

Note: We need a qualification to the assumption that A.C.E. accurately determined household population duplicates within the search area. Tom Mule has pointed out that while the A.C.E. production matching attempted to find all such duplicates within the sample block cluster, when it looked for duplicates in the surrounding blocks it looked only at persons in housing units identified as geocoded to the surrounding blocks and found to be duplicated there (Childers 2000, p. 39 and pp. 45-46). So while the A.C.E. may have accurately determined those duplicates that it looked for, it intentionally did not look at everyone in the surrounding blocks for duplicates. Fay (2002, p. 16) notes that Mule (2001) found 146,880 duplicates in surrounding blocks by computer matching whereas the March 2001 A.C.E. estimate was only 98,335. Since A.C.E. found significantly more duplicates within the sample block clusters than did the computer matching, some of the difference in the surrounding blocks is presumably due to A.C.E. not attempting to find all the duplicates there. Fay (2002, p. 33, footnote (3)) also noted that further analysis would be required to determine if the Measurement Error Review adequately accounted for the surrounding block duplicates. Thus, there remain questions about how to best account for the duplicates in surrounding blocks in tabulations of CEs. (The Targeted Extended Search (TES) operation collected data in surrounding blocks only for a sample of the A.C.E. block clusters, but this sampling is accounted for by the TES sampling weights, and so does not cause the sort of problems just mentioned.)

References

Bean, Susanne L. (2001), "ESCAP II: Accuracy and Coverage Evaluation Matching Error," Executive Steering Committee for A.C.E. Policy II, Report 7, October 12, 2001.

Childers, Danny R. (2000), "Accuracy and Coverage Evaluation: The Design Document," DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1, October 11, 2000.

Fay, Robert E. (2002), "ESCAP II: Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee for A.C.E. Policy II, Report 9, Revised Version, March 27, 2002.

Feldpausch, Roxanne (2001), "ESCAP II: E-Sample Erroneous Enumerations," Executive Steering Committee for A.C.E. Policy II, Report 5, October 14, 2001.

Fenstermaker, Deborah and Davis, Peter (2002), "A.C.E. Revision II: Estimated Correct Enumeration and Residence Probability for Duplicate Links," DSSD A.C.E. Revision II Memorandum Series #PP-52, Decennial Statistical Studies Division, U.S. Bureau of the Census, December 31, 2002.

Jones, John (2003), "Person Duplication in the Search Area Measured by the 2000 Accuracy and Coverage Evaluation," Census 2000 Evaluation O.16, Decennial Statistical Studies Division, U.S. Bureau of the Census, forthcoming.

Kostanich, Donna L. (2003), "A.C.E. Revision II: Design and Methodology," DSSD A.C.E. Revision II Memorandum Series #PP-30, Decennial Statistical Studies Division, U.S. Bureau of the Census, March 11, 2003.

Martin, Elizabeth (2001), "Instrument Differences and their Possible Effects: Comparison of the Evaluation Followup (EFU) and Person Followup (PFU) Instruments," internal Census Bureau note, October 12, 2001.

Mule, Thomas (2001), "ESCAP II: Person Duplication in Census 2000," Executive Steering Committee for A.C.E. Policy II, Report 20, October 11, 2001.

Mule Thomas (2002), "A.C.E. Revision II: Further Study of Person Duplication," DSSD A.C.E. Revision II Memorandum Series #PP-51, Decennial Statistical Studies Division, U.S. Bureau of the Census, December 31, 2002.