Estimating Erroneous Enumeration in the US Decennial Census using Four Lists

G. Gordon Brown, RTI International, Paul P. Biemer, RTI International and University of North Carolina
and Dean H. Judson, Bureau of the Census

Keywords: dual system estimation; census undercount; latent class analysis; capture-recapture

Introduction

Erroneous enumerations occur when non-residents encountered in the census enumeration processes are erroneously counted as residents. Erroneous enumerations include persons who were deceased prior to Census Day, born after Census Day, or otherwise not residents of the target area on Census Day. They also include geocoding errors, duplicated persons, and fictitious or nonexistent persons. We will refer to all of these entities as nonresidents regardless of their source. Any nonresident who is classified as a resident in the enumeration process will be called an erroneous enumeration (EE).

The existence of EEs in the counting process has long been a concern for population censuses. Adjustments of the estimates of the population total count, *N*, for EEs is an essential component of the census undercount evaluation process and has been used for many years. In the U.S. Decennial Census, an estimate of the number of EEs in the census is provided by a component of the post-enumeration survey (PES) called the E-sample which is essentially a sample of persons that were counted in the census (see, for example, Hogan, 1993). However, for Census 2000, a substantial number of EEs were not identified by the E-sample survey and were included in the estimates of total population counts thereby inducing bias in the estimates of census coverage error (ESCAP, 2001).

Biemer, Brown, Judson and Wiesen (2001) and Biemer, Brown, and Wiesen (2004) develop an approach that allows for varying levels of undetected EEs in the population lists using a type of finite mixture or latent class model. Their approach requires a third enumeration or listing of the population, which in our application, is derived from merging records from administrative systems (Judson, 2000). The resulting data take the form of an incomplete contingency table with an unobserved fourth dichotomous variable representing an individual's true status (i.e., population member or non-population member). Biemer, et al. (2001) described how to employ latent class models to estimate the expected values of the observed cells of this table and then to project these estimate onto the unobserved cells in order to estimate the total number of population members. This paper is an extension of the work presented in Biemer, et al (2001) and in Biemer, et al (2004). To accommodate the page limitations imposed on this paper, we assume the reader is familiar with the notation and methodology used in the referenced papers.

Biemer, et al (2004) discusses the identifiability of the latent class model (LCM) denoted by L_{tAB}, which, in the notation of hierarchical log-linear models is {AX, BX, CX, AB), as well as other L-models containing interactions. They show that the L-models such as this one that involve list-by-list interaction terms are not identified unless the erroneous enumeration rates, γ's, are known for each list containing undetected EE's. One solution to this dilemma they consider is to estimate the γ's by fielding random samples of cases from each list and reinterviewing the sample members to determine their true Census Day residential status. The estimate of the proportion of cases on a particular list that are classified as residents but are non-residents provides an estimate of γ for that list.

This method of obtaining the γ's is costly and could result in estimates of the parameters which are still considerably biased due the inability to accurately determine whether census enumerations and administratively records list entries are Census Day residents. However, the work of Biemer, Woltman, Raglin, and Hill (2000) may offer a means for estimating the γ's with little or no additional fieldwork using a quadruple system model. In this paper, we explore the feasibility of using this approach for providing identifying parameters to the L-models for the estimation of *N*.

Notation

Define the latent variable *X* as in Biemer, et al. (2001); i.e., X=1 denotes a resident and $X = 2$ a nonresident. Define the indicator variable *A* for each individual in the Census as follows:

$A = 1 \Rightarrow$ the individual is rostered in the census

$A = 2 \Rightarrow$ the individual is not rostered in the census.

Define the indicator variable *B* for the PES as follows:

B' = 1 ⇒ the individual is observed in the PES and is classified as a resident

B' = 2 ⇒ the individual is observed in the PES and is classified as a nonresident

B' = 3 ⇒ the individual is unobserved in the PES (i.e., not rostered).

Define C' analogously for the measurement error reinterview (MER) classification. The MER was a component of the Census 2000 process and is essentially a reinterview of a sample of PES persons for the purpose of evaluating the PES classification accuracy and other measurement errors associated with the PES. Finally, define D for the administrative records list as follows:

D = 1 ⇒ if the individual is listed on the administrative records list (i.e., observed and classified as a resident)

D = 2 ⇒ if the individual is not on the list (unobserved)

These four lists are combined to form the AB'C'D table. It should be noted in some cases it may be advantageous to define three categories for D; i.e, (1) on the list and deemed to be a resident, (2) on the list and deemed to be a nonresident, and (3) not on the list. This is discussed in a later section of this paper.

There are a number of structural zeros in the $AB'C'D$ table. For example, we assume that if C = 3, then A = 2 and B = 3 with certainty in order to conform to the Census 2000 procedures. These constraints arise because once an individual (either resident or non-resident) has been identified in either the census or in the PES, that individual is automatically carried forward to the MER for verification purposes. Similarly, we assume that if B = 3, then A = 2 with certainty; i.e., census persons are carried forward automatically to the PES for verification. It can easily be shown that the $AB'C'D$ table contains 36 cells and 15 structural 0's for a total of 21 non-zero cells.

A possible L-model that can be fitted to these data is the $L^{AB'C'D}_{tAB'.B'C}$ model with terms listed in Table 1. The model in Table 1 is essentially an extension of the L_{tAB} to four systems. As in our previous work, we assume the error in the administrative records list, now denoted by D, is locally independent of the errors in the other lists (i.e., no AD, $B'D$, or $C'D$ interactions). However, interactions are introduced among the other three lists - i.e., AB' or $B'C'$. This model is not identified in general (see Biemer, et al, 2004), but is identified if γ_A, γ_B, γ_C, and γ_D are known and specified in the model through the appropriate constraints. The parameters γ_A, γ_B, γ_C can be estimated through additional fieldwork as noted above. However, an alternative, two-stage estimation approach is described in the next section that eliminates the need for additional fieldwork.

Table 1. The $L^{AB'C'D}_{tAB'.B'C}$ Model for Four Systems

Conditional Probability	Submodel[1]	D.F.
X	$\{X\}$	1
$A\|X$	$\{AX\}$	2
$B'\|AX$	$\{B'X \ AB'\}$	6
$C'\|AB'X$	$\{C'X \ B'C'\}$	8
$D\|ABCX$	$\{DX\}$	2
Total		19

[1]Standard hierarchical log-linear modeling notation is used.

Two-Stage Estimation Approach

We consider a two-stage approach for fitting models to the full $AB'C'D$ table. In the first step, we fit a model to a collapsed version of table $AB'C'D$ denoted by $ABCD$ to produce estimates of γ_A, γ_B, γ_C, and γ_D. Collapsing the table prior to fitting the model eliminates the structural zeros and simplifies the model specification. Under fairly general assumptions, models for this collapsed table which specify unknown γ_A, γ_B, γ_C, and γ_D are identifiable.

Once estimates of γ_A, γ_B, γ_C, and γ_D are obtained, they can be treated as known constants in the models for the full table $AB'C'D$ in order to yield an estimate of N, the total population size adjusted for census misses and EE's in all systems. In addition, the estimates for the γ's can be used for either dual or triple system models since the γ's are specific to the lists used, not the subsequent models in which they are employed.

To obtain estimates of γ's, we consider a model based upon the $AB'C'D$ tables collapsed to the $ABCD$ table where

B = 1 ⇒ the individual is observed in the PES and is classified as a resident (same as B' = 1)

B = 2 ⇒ the individual is observed in the PES and is classified as a nonresident or the individual is unobserved in the PES (i.e., not rostered).

Note that B = 2 combines B'= 2 and B' = 3. The

collapsed indicator *C* is defined analogously to *B* for the MER. Table 2 shows the twelve cells from the *AB'C'D* that are collapsed and to the four cells they are collapsed into for the *ABCD* table. The result is that the *ABCD* table contains 16 cells with no structural zeroes. Table 2 summarizes the cells that are collapsed to form the *ABCD* table.

The collapsed *ABCD* table allows us to fit more complex models due to the removal of the structural zero in the unobservable cell. For example, if we fit the model in Table 3 to these data, we can obtain estimates of γ_A, γ_B, γ_C, and γ_D which can then be provided to the model in Table 1, or any Triple System or Dual System model, to obtain an identifiable model. Fitting a model to the collapsed table can thus be regarded as the first stage of a two-stage estimation process.

Table 2. Collapsed cells from *AB'C'D* to *ABCD*

AB'C'D	*ABCD*
1221 1321 1331	1221
2221 2321 2331	2221
1222 1322 1332	1222
2221 2322 2332	2222

To see this, note that P(*B*=1|*X*=2) for the $L^{ABCD}_{tAB.BC}$ model in Table 3 is equal to P(*B'*=1|*X*=2) for the model $L^{AB'C'D}_{tAB'.B'C}$ in Table 1. Likewise, P(*C*=1|*X*=2) = P(*C'*=1|*X*=2). Thus, we can estimate P(*A*=1|*X*=2), P(*B*=1|*X*=2), P(*C*=1|*X*=2) and P(*D*=1|*X*=2) or, equivalently, γ_A, γ_B, γ_C, and γ_D, using the $L^{ABCD}_{tAB.BC}$ model or a similar L-model (see Table 4). These estimates can then be supplied to any Quad, Triple or Dual System model to produce estimates of *N*.

Identifiability of the First Stage Model

A number of other models can be specified for the Quad-system estimator (QSE). Some of these are listed in Table 4. Note, however, that a number of models that would be of interest are not identifiable. By necessity we will confine our study to identifiable QSE models, specifically the one defined in Table 3. As seen in Table 4, there are still a number of identifiable models which are still useful for the undercount evaluation application.

Table 3 The $L^{ABCD}_{tAB.BC}$ Model for Four Systems

Conditional Probability	Sub-model	D.F.
X	{*X*}	1
A|*X*	{*AX*}	2
B|*AX*	{*BX AB*}	3
C|*ABX*	{*CX BC*}	4
D|*ABCX*	{*DX*}	2
Total		12

Simulating Triple and Quadruple System Populations

In order to determine the feasibility of using a QSE model to estimate the γ's, we need to generate a series of populations with known parameters, apply the two-stage process to these populations, and then compare the model estimates with the known parameter estimates. The expected value method (which we referred to as "expectulation") used in our earlier work was used to generate artificial samples. For this method, data for any specified sample size are generated that exactly correspond to the expected cell counts for the model assumed for the artificial population. Thus, the model being evaluated is fit to cell counts that correspond exactly to the assumed population model. We showed in Biemer, et al (2001) that Monte Carlo simulation and the expectulation method tended to yield nearly identical inference for the models we investigated. We will continue to proceed with our analysis of the new models using this approach. Brief descriptions of both methods are given in the following paragraphs.

For Monte Carlo simulation, multiple samples, say 1000, are generated from an artificial population. The model of interest is then fit to the all of the 1000 artificial samples and the point estimates for the parameters for each of the 1000 simulations is recorded. The estimated value for any parameter can then be calculated by taking the average of the 1000 point estimates from the artificial samples.

For the expectulation method, the expected number of observations in each cell of the *ABCD* table is calculated using the population parameters from the artificial population. The model of interest is then fitted to this 'sample' of expected values. The point estimates for the parameters of interest are then recorded from this single model fitting. The advantage of expectulation is that point estimates are easier to obtain since the models only needed to be fit to a single data set.

Table 4. Some Other QSE Models and Their Identifiability

Model	Specification of Conditional Terms				Identified?
	AX	*B|AX*	*C|ABX*	*D|ABCX*	
1	*AX*	*BX AB*	*CX CA CB*	*DX*	Yes
2	*AX*	*BX AB*	*CX CA CB*	*DX DA*	No
3	*AX*	*BX AB*	*CX CA CB*	*DX DB*	No
4	*AX*	*BX AB*	*CX CA CB*	*DX DC*	No
5	*AX*	*BX AB*	*CX*	*DX*	Yes
6	*AX*	*BX AB*	*CXB CA*	*DX*	No
7	*AX*	*BX AB*	*CXA CB*	*DX*	Yes
8	*AX*	*BX AB*	*CX CB*	*DX DA*	Yes
9	*AX*	*BX AB*	*CX CB*	*DX DB*	No
10	*AX*	*BX AB*	*CX CB*	*DX DC*	No

The ℓEM package (Vermunt, 1997) was employed to fit the models to the data generated from the artificial population since it is able to fit this model precisely. The accuracy of the ℓEM estimate of γ_D over many artificial populations, varying γ_D as well as the other parameters, is the criteria we used for determining the feasibility of the estimating γ_D .

In our previous work (Biemer, et al 2001), we considered the accuracy of estimating *N* using the L_{tAB} model for various levels of error in γ_D. We found that reasonable estimates of γ_D would yield estimates of *N* that are viable for census coverage evaluation. For this reason, it is not necessary to fit the model in Table 1 to determine the feasibility of the two-stage approach. Essentially, if the error in γ_D is small, the accuracy of estimates of *N* will be adequate and the two-stage approach is feasible.

Other experiments were conducted with more complex populations as well. For example, we examined the approach for populations generated by the models L_{tAB}^{AD}, L_{tAB}^{ABD}, and L_{tAB}^{ABCD} as well as other combinations of lists with EE's. In addition, we considered populations that incorporate the *BC* interaction as well as erroneous enumerations in one or more lists.

Results of the Feasibility Study

The results and discussion presented in this section should be regarded as preliminary. While most of our conclusions and remarks are supported by data and theory, some have the flavor of conjecture which is based on our knowledge of LCMs, capture-recapture models, and the census. The goal of this investigation is not to provide definitive answers, but rather to alert the reader to a potentially powerful tool in estimation of census undercount. More work is needed to fully address all of the major issues related to the QSE methodology.

Modeling Assumptions

Several important assumptions are made for the models presented in this paper:

1) There are only two types of individuals in the population and they are defined as the residents and non-residents.
2) Heterogeneity of enumeration probabilities for residents is adequately modeled by the grouping variables (race, gender, age, etc.) in the model.
3) Likewise, heterogeneity of enumeration probabilities for non-residents is adequately modeled by the grouping variables in the model.
4) On any list, the enumeration probability for the residents is not equal to the enumeration probability for non-residents within the cells defined by the grouping variables in the model.
5) The probability that an individual in the population is missed by all four lists is quite small.

These assumptions guarantee that there are two distinct types of individuals in the population having unique sets of enumeration probabilities. In the following, we confine our analysis to a single group in the population. However, extension to multiple groups defined by the cross-classification of grouping variables is straightforward.

Latent class analysis distinguishes between residents and non-residents in the population by their differential probabilities of enumeration. In general, the classification of individuals into the two types will be more accurate with greater differences in enumeration probabilities between residents and nonresidents. Since all of the models we use in the present paper have two latent classes, under assumptions (1)-(5), we can expect that one of the latent classes will represent residents and the other will represent

non-residents. In most cases, it is not difficult to identify which latent class represents which type of individual. Since we expect residents to outnumber non-residents, the class with the largest estimated population size is usually the resident class.

In addition, we expect that non-residents will have lower enumeration probabilities than residents. Therefore, the class having the higher enumeration probabilities will usually represent the resident class. Once this determination is made, the EE probabilities and γ's are can be computed based upon the misclassification probabilities for the non-resident class. Of course, in some areas of the country, these two criteria will not be successful in identifying residents and nonresidents primarily as a result of the failure of at least one of the five assumptions to hold.

For example, as noted above, an important assumption is that the enumeration probability for residents and nonresidents be different on any given list. In the worst case, residents and non-residents have the same enumerations probabilities on all of the lists. In this situation, our models could not recognize two distinct groups and would breakdown. This is a violation of the 4th modeling assumption.

A second example is that in some areas of the countries there may be two groups of resident some have small enumeration probabilities - so-called "Hard to Enumerate (HTE)" individuals - and the remaining have large enumeration probabilities. In this situation, the L-models will incorrectly classify the HTE residents as nonresidents. As a result, the estimates of γ will be biased upward and estimates of *N* based on QSE parameters will be biased downward. The reverse situation may also be true. Suppose there are two groups of non-residents with one group having a set of enumeration probabilities approximately equal to those for residents. In this situation, the EE rates will be underestimated and the bias in the estimate of *N* will tend to be positive. These are examples of violations of modeling assumptions (1) and (2).

To further generalize the above examples, it is possible that the enumeration probabilities for residents come from a distribution, call it *R*, and the enumeration probabilities for the non-residents comes from a different distribution, say *NR*. If the two distributions are widely separated (i.e., the means of the two distribution are very different), then the heterogeneity of enumeration probabilities will have little effect on the L-models' ability to recognize residents and non-resident. However, if the two distributions overlap considerably (i.e., very little separation between means of the distributions), then the L-models will produce poor quality inference with large biases in the parameters of interest. Therefore, it appears that the amount of heterogeneity in the enumeration probabilities that can be tolerated by the L-models depends to a large extent on the degree of separation between resident and nonresident enumeration probabilities.

Another key assumption of our approach is that every resident appears on at least one of the lists as either a resident or nonresident. More specifically, we assume that it is highly unlikely that residents are not listed in some way at least on one of the four lists. This is necessary since we assume that the (2,3,3,2) cell (i.e., the cell representing residents missed by all four systems) is an observed zero. In reality, the (2,3,3,2) cell in the *AB'C'D* table is unobservable and should be treated as a structural zero as was done in Biemer, et al (2001) and Biemer, et al. (2004). However, that approach is not possible with the current models in order to achieve identifiability.

Although this assumption is not true in general, it may still be reasonable if the probability of an observation falling into the (2,3,3,2) cell is small. With four lists, it is likely that any individual will be listed on at least one of the lists if the enumeration probabilities for each list are moderate to large.

Estimation of EEs for Populations Following the $L_{tAB.BC}^{ABCD}$ Model

The focus of our investigation for this paper is the model $L_{tAB.BC}^{ABCD}$. We used the $L_{tAB'.B'C'}^{AB'C'D}$ model to generate an artificial population, collapsed this population to the *ABCD* table, and used the $L_{tAB.BC}^{ABCD}$ model to estimate the parameters of interest from that population. As noted previously, our primary interest centers on the estimates of the γ's. Since our main objective is to assess the feasibility of the two-stage estimation approach, we examined the most favorable case conceivable where the estimation model and population model virtually agree. If the estimation model performs poorly under this best case scenario, it should not be expected to perform well under more adverse modeling situations. On the other hand, if the two-stage approach works well under the most favorable conditions, exploring the approach

under situations where the population and estimation models differ systematically would be justified.

The $L^{ABCD}_{tAB.BC}$ model contains an *AB* interaction, a *BC* interaction, and allows erroneous enumerations on all four lists. We assume that EEs enter the sample with the same structure as the residents, but have a different set of enumerations probabilities. By virtue of the absence of the *ABX* and *BCX* interaction terms, we tacitly assume that *AB* and *BC* interactions do not depend on the latent class (residents and nonresidents). The *AB* and *BC* interactions are needed due to the level of dependence of the PES on the Census and the MER on the PES. The interactions are also capable of detecting levels of heterogeneity within the lists and potential behavioral effects due to interaction with the census takers. This model is not the most complex model allowable given the degrees of freedom available for modeling, but it is, perhaps, the simplest model that is still plausible under actual census conditions. More complex, identifiable models are listed in Table 3, but are not explored in this paper.

Given the exploratory nature of the current investigation, the main focus of work to date has been on the estimation of $\gamma_{D.}$ Although $\gamma_{A.}$, $\gamma_{B.}$, and γ_C can also be estimated from the model, we have not fully investigated the accuracy of the estimates of these other parameters. Estimation of γ_D (as well as the other γ's) for this model is quite straightforward using ℓEM. Although ℓEM does not directly calculate the EE rates, the ℓEM output provides estimates of the enumeration probabilities for the resident and nonresidents, in particular, P(*D*=1|*X*=1) and P(*D*=1|*X*=2). These probabilities can be used to calculate γ_D using the following formula:

$$\gamma_D = \frac{P(D=1 \mid X=2)}{P(D=1 \mid X=2)+P(D=1 \mid X=1)}.$$

For the $L^{ABCD}_{tAB.BC}$ model, an interesting result is that γ_D tends to be underestimated when the observed cell in the *AB'C'D* table (i.e., the (2,3,3,2) cell) is treated as an observed 0 rather than a structural 0. When the probability of a resident in the (2,3,3,2) cell, denoted by P(2,3,3,2|*X*=1), is quite small, the estimates of γ_D are essentially unbiased. As P(2,3,3,2|*X*=1) increases, estimates of γ_D tend to be negatively biased.

Table 5 illustrates several examples of where P(2,3,3,2|*X*=1) is relatively small and the estimates of γ_D from the model $L^{ABCD}_{tAB.BC}$ are virtually unbiased. In this table, the expected number of residents in the (2,3,3,2) cell is denoted by P(2,3,3,2|*X*=1)×*N* and is given in column 2. The total number of individual in the population was 10,000 and P(*X*=1) in column 1 shows the probability of an individual being a resident. The estimate for γ_D denoted by $\hat{\gamma}_D$ and the corresponding true value, $\gamma_{D,}$ are given in the columns 3 and 4, respectively.

Table 5. Estimates of γ_D for the $L^{ABCD}_{tAB.BC}$ Model with P(2,3,3,2) Small

P(*X*=1)	P(2332|*X*=1)×*N*	$\hat{\gamma}_D \times 100$	$\gamma_D \times 100$
0.7	0.08	6.75	6.67
0.8	0.42	2.81	2.93
0.8	0.55	2.92	2.93

In Table 6, the proportion of nonresidents entering the sample is increased. In this scenario, the probability P(*D*=1|*X*=2) is set to 0.20 insuring that there will be a large number of EEs on the *D*-list. As with Table 6, the P(2,3,3,2|X=1) remains small. The estimated values of γ_D is given and compared to the true value. While the actual bias in the estimate of γ_D tends to increase as the number of non-residents increase, the percent bias tends to remain relatively constant. This indicates that our method of estimating the γ's will work when the proportion of nonresidents in the sample is large.

Table 6. Error in γ_D When Nonresidents Are Increased

P(*X*=1)	P(2332|*X*=1)×*N*	$\hat{\gamma}_D$ x100	$\gamma_D \times 100$
0.7	2.8	12.8	13.9
0.8	1.8	8.1	8.7
0.8	3.2	7.8	8.6
0.9	3.6	3.6	4.0

Extension of *D*-list to 3 levels: the *D*'-scheme

One potential problem is that the number of correctly identified non-residents is not available from the *D*-list. It is possible that this piece of information could improve the quality of the

estimates of the γ's. For example, assume that a given person is correctly determined to be a non-resident on the *D*-list. Further assume that this person was incorrectly determined to be a resident on the *A*-, *B*- or *C*- list. Under the current scheme, the information from the *D*-list would not be recorded and the person would be incorrectly identified as a resident. Clearly, this would bias the estimates of γ.

In order to account for the situation described above, we propose a three level scheme for the ARL which we will refer to as the *D*'-scheme. For the *D*'-scheme, there would be two stages in the formation of the ARL. Stage 1 would consist of a preliminary compilation of the lists used to form a pre-ARL list. This list would contain N_{pre} entries. Stage 2 would consist of breaking the pre-ARL list into three parts as follows:

D'=1 the individual is on the pre-ARL and classified by the ARL as a resident

D'=2 the individual is on the pre-ARL and classified by the ARL as a non-resident

D'=3 the individual is unobserved by the pre-ARL list

The next step would be to build the *AB*'*C*'*D*' table and collapse this to the *ABCD* table in a similar fashion as was done with the *AB*'*C*'*D* table. In this manner the more information regarding non-residents in the population is known thereby, by conjecture, improving the estimates of the enumeration probabilities.

Initial simulations using *D*' rather than *D* showed promise. Specifically, as the number of nonresident on the pre-ARL were located, correctly identified and placed in the *D*'=2 cell, estimates for γ_D improved. For example, the model used to generate the data in Table 7 had a high probability of a resident being in the (2,3,3,2) cell which resulted in a large bias in γ_D. Estimates of γ_D from the data generated by the *D*'-scheme model showed less bias when compared to the *D*-scheme. Specifically, the data set used in Table 7 had 112.4 non-residents in the (2,3,3,2) cell under the *D*-scheme. Under the *D*'-scheme, some of these non-residents are located and correctly identified as non-residents. In one simulation, more than half of the non-residents were correctly identified resulting in an estimate of γ_D of 0.067. While still bias, this is an improvement over the estimate of 0.044 from the *D*-scheme.

The reasons for the improvement are largely based on conjecture. The ARL or *D*-list should act as an independent assessment of individuals available for rostering on the census. It is likely that non-residents that are incorrectly classified as residents by the census are correctly classified as non-residents by the ARL. If these correctly classified non-residents can be successfully linked to their census records, we gain more information regarding the inclusion of EEs on all of the lists be discussed. If individuals classified as non-residents are discarded by the ARL as is currently done, then this potential information is lost and cannot be regained by any of the models that we have explored.

Table 7. Error in γ_D comparing *D*- and *D*'-scheme

P(*X*=1)	Scheme	P(2332|*X*=2) ×*n*	$\hat{\gamma}_D$ ×100	γ_D ×100
0.8	*D*	112.4	4.4	9.5
0.8	*D'*	30.32	6.7	9.5

Conclusions and Discussion

We believe that the QSE approach proposed here has the potential to be useful in estimation of EEs in the census enumeration system. However, considerably more work needs to be undertaken in order to better perfect the methods and models presented. Several key issues regarding these models should be further explored. Most notable among these is the robustness to heterogeneity in the enumeration probabilities for the residents. Full exploration of all of the assumptions should be undertaken to determine when QSE models can be used successfully. Nevertheless, the models clearly show the potential to estimate all of the parameters necessary to obtain quality inference regarding the erroneous enumeration rates and, therefore, a quality estimate of the census day population size.

Some advantages of this approach are: a) estimates of erroneous enumeration rates can be calculated with currently available data, (b) no additional field work would be required, and (c) the bias in these estimates are minimal provided that all of the assumptions specified are met.

One disadvantage is that the assumptions may be too restrictive in some census areas; however, it may be possible to relax some of the restriction given further research. Also, the models may not be robust to some of the stated assumptions. With the exceptions noted in the paper, we have not studied the robustness of these models to all of the stated assumptions.

Finally, the models can be difficult to fit. Even using a latent class software package like ℓEM, fitting there models can be challenging.

The primary focus of the work shown here was to estimate the EEs on the ARL or *D*-list. Using these models it is also possible to estimate EEs on either the PES or *B*-list, which is used to compose the Dual System Estimates currently being explored by the Census Bureau. Using these models to estimate other parameters, such as the *AB* interaction or $\gamma_{B,}$ is possible given the same set of assumptions used for γ_D. The exact effect of the violation of these assumptions on the parameters specified has not been explored.

References

Biemer, P. P., Brown, G. G., Judson, D. H., and Wiesen, C. (2001). "Triple System Estimation with Erroneous Enumerations in the Administrative Records List." *2001 Proceedings of the American Statistical Association*, Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association.

Biemer P.P., Brown G.G. and Judson D.H. (2004). Latent Class Models for Evaluating the Accuracy of Census Counts. *2004 Proceedings of the American Statistical Association*, Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association.

Biemer, P., Woltman, H., Raglin, D., and Hill, J. (2001) "Enumeration Accuracy in a Population Census: An Evaluation Using Latent Class Analysis," *Journal of Official Statistics*, Vol. 17, No. 1, pp. 129-148

ESCAP. 2001. ESCAP II Report No. 1 Recommendation and Report of the Executive Steering Committee for A.C.E. Policy (ESCAP II), Bureau of the Census, Washington, D.C.

Hogan, H. (1993). "The 1990 Post-Enumeration Survey: Operations and Results," *Journal of the American Statistical Association*, Vol. 88, No. 423, 1047-1071.

Judson, D. H. 2000. "The Statistical Administrative Records System: System Design, Successes, and Challenges." Internal Census Bureau Report, Nov. 11, 2000.

Vermunt, J. 1997. *LEM: A General Program for the Analysis of Categorical Data*, Tilburg, University.