Journal of Official Statistics, Vol. 17, No. 1, 2001, pp. 129–148

Enumeration Accuracy in a Population Census: An Evaluation Using Latent Class Analysis

Paul P. Biemer[1], Henry Woltmann[2], David Raglin[2] and Joan Hill[2]

To evaluate the coverage error in a population census enumeration, many countries conduct a Post Enumeration Survey (PES) which is designed to identify individuals who were missed in the Census or individuals who were counted that should not have been. The quality of the PES evaluation of coverage error is only as good as the quality of the PES itself and much effort has been devoted world-wide to improving the PES enumeration methodology. In this article, we apply latent class analysis (LCA) to evaluate the quality of the PES and compare estimates of the PES classification error with the corresponding estimates from a traditional analysis of these data. The primary basis for these evaluations is a reconciled reinterview survey of the PES respondents. The traditional analysis treats the reconciled reinterview survey results as infallible and attributes disagreements between the PES and the reconciled reinterview classifications to deficiencies in the PES. With LCA, the reinterview results are treated as fallible measures which simply produce another indicator of true residence status. LCA estimates of the error probabilities for all three classifiers (the Census, the PES, and the reconciled PES reinterview) are obtained by maximum likelihood estimation under an assumed latent class model. In this article, we demonstrate the use of LCA for evaluating post enumeration survey accuracy and summarize the key findings for PES evaluation studies.

Key words: Coverage error; reinterview; census undercount; integrated coverage measurement; nonsampling error; classification error.

1. Introduction

Obtaining an accurate count in population censuses has been a concern in most countries throughout the world. In the United States, errors in the population counts, particularly undercoverage errors, have been well-documented in censuses since the 1940s. The U.S. Census Bureau as well as its counterparts in a number of other countries have relied primarily on a postenumeration survey (PES) methodology to evaluate the census undercount (see, for example, Brown, Diamond, Chambers, and Buckner 1999; Hogan 1993; Choi, Steel, and Skinner 1988). A PES is a sample survey conducted after the Census for the purpose of enumerating individuals in the sample households. By attempting to match individuals enumerated in the PES to individuals enumerated in the Census, it is possible to obtain data to estimate the proportion of the true population that was missed in the Census as well as the number of erroneous enumerations (EE's) in the Census (e.g., duplicated people, fabricated households, people who died or moved out of the area before Census day). In the estimation process, persons in the PES are cross-classified by whether

[1] Research Triangle Institute (RTI), PO Box 12194, Research Triangle Park, NC 27709-2194, U.S.A. E-mail: ppb@rti.org

[2] U.S. Census Bureau, Washington, DC 20233, U.S.A.

they were counted in the original Census or not to form a 2×2 table with one empty cell corresponding to individuals missed in both the PES and the Census. To complete the table, dual system estimation (Chandrasekar and Deming 1949; Wolter 1986) is used. Using this method, a dual system estimator (DSE) of the census coverage error can be derived for virtually any geographic area of the country.

It is well-known that the PES and the PES-Census matching process is also subject to errors that arise from incorrect matching, tracing difficulties, and many of the same problems that cause Census error (see, for example, Hogan 1993). In addition, the dual system estimator of the total population is subject to a number of additional biases arising from heterogeneous enumeration probabilities in the population and the lack of independence in enumeration error between the two systems (Wolter 1986). For the 1990 Census, the U.S. Census Bureau mounted a considerable research program to examine the major sources of error in the dual system estimator in preparation for the 1990 Census. Some of the results of this research are reported in Mulry and Spencer (1993). This research determined that the largest contributor of bias in estimates census undercount is PES enumeration error which are errors in the number of individuals counted for a particular PES household. Consequently, considerable emphasis was placed on improving the quality of the PES in preparation for Census 2000.

Traditional methods for evaluating census and survey data collection operations have relied on replication with reconciliation (see, for example, U.S. Bureau of the Census 1985; Mulry and Spencer 1993; Kuha and Skinner 1997). Using these methods, an operation to be evaluated is repeated, often by operators having higher skill levels than the original operators, and any differences between the results of the initial operation and the replication are reconciled in order to arrive at the best possible results. The results of the reconciled replication process are then used as a ''gold standard'' and treated as infallible for purposes of evaluating the results of the initial operation.

Applying this method to the evaluation of the PES, the reconciled replication method takes the form of a reconciled reinterview. A household respondent living at a PES address is reinterviewed soon after the PES interview for the purpose of obtaining a third roster of all household members living at the address on Census Day. Since this third roster is obtained without reference to either the Census questionnaire roster of the PES roster, it is referred to as the ''independent'' Census Day roster. The independent evaluation roster is then compared with two previously obtained rosters and any discrepancies between these lists are reconciled with the reinterview respondent. The result is a final, ''gold standard'' roster which is used to identify enumeration errors in both Census and PES rosters.

An important limitation of the gold standard methodology is the assumption that the reconciled reinterview yields the truth. Studies examining the validity of this assumption in other contexts have shown that the reconciled reinterview approach is subject to considerable measurement errors (see, for example, Biemer and Forsman 1992; Sinclair and Gastwirth 1993). For example, there is evidence that reinterview respondents satisfice[3] during the reconciliation process and that the reinterviewers often fail to follow

[3] ''Satisfice'' is a term which means that the respondent exerts minimal cognitive effort in responding to a survey question. These are often ''top of the head'' responses that are prone to measurement error (see Krosnick and Alwin 1987).

reinterview procedures (see Biemer and Hubbard 1996). Further, little is known about the cognitive processes operating during reconciliation and there is no well-established cognitive theory to guide the design of reconciliation procedures. Thus, using the reconciled reinterview results as truth can result in misleading conclusions regarding the quality of the PES. This point was well-noted in West and Griffiths (1996) and Raglin, Griffin, and Kromar (1998), which applied traditional methods to the 1995 and 1996 PES evaluations.

Latent class analysis (LCA) is an alternative methodology for quality evaluations that also uses replicate measures to assess the measurement error in survey results. However, LCA does not require the assumption that one of the replicate measures is the truth or a gold standard. Rather, all of the indicators in the analysis are assumed to be fallible, with either correlated or uncorrelated errors between the measurements. The analyst posits a model for the measurement error distribution associated with each measurement process and tests the models against the observed data using conventional chi-squared goodness-of-fit criteria. The best latent class model is one which is plausible, parsimonious, and fits the data well. The best model is then used to generate estimates of the classification error probabilities for the measurement processes under investigation.

In 1995 and 1996, the U.S. Census Bureau conducted two tests of the PES procedures for a program referred to as the Integrated Coverage Measurement (ICM) program. In two previous reports (West and Griffiths 1996; Raglin, Griffin, and Kromar 1998), enumeration accuracy of the PES was evaluated using the traditional gold standard reinterview approach. In this article, we reanalyze these data using latent class analysis (LCA) approaches and the results are compared with those of the analyses of West et al. and Raglin et al. The goal of the article is to investigate the potential of the LCA methodology for PES evaluation as an alternative or supplement to traditional gold standard analysis.

In the next section, we briefly describe the PES designs for the test censuses as well as the corresponding reconciled reinterview survey designs used to evaluate the PES enumeration error. Section 3 contains a brief description of the LCA methodology that was implemented in the study. In Section 4, the LCA methodology is applied to each year's data and the basic error parameters are estimated and compared with those from the traditional analyses. Finally, in Section 5, we summarize our findings and provide our recommendations for the Census 2000 evaluation.

2. The 1995 and 1996 U.S. Test Census-PES Evaluation Studies

The 1995 and 1996 Test Censuses were conducted by the U.S. Bureau of the Census to test and evaluate a number of alternative data collection and estimation methodologies that were being considered for the Census 2000. For each test census, a PES was conducted following the census and the enumeration error in these surveys was evaluated using traditional reconciled reinterview methods. Thus, the reinterview surveys were designed to obtain the ''true'' residence status classifications for all persons in the reinterview households. Some details of the PES and the PES evaluation reinterview surveys for both years are given below.

2.1. PES design

In both the 1995 and 1996 Test Censuses, the Census questionnaires were completed either by the respondent through a self-administered census questionnaire or by an enumerator interviewing the respondent using paper and pencil interviewing methods during the Nonresponse Followup operations. The 1995 PES was conducted a few months after Census Day in Oakland, California, using Computer Assisted Personal Interviewing (CAPI). PES interviewers visited each household to obtain a new listing of all individuals living at a sample address on Census Day and then compared this list to the list of individuals from the Census questionnaire, referred to as the Census roster. People appearing on both rosters were linked and considered to be correctly enumerated. For those that did not match, the PES respondent was asked to provide information on the living situations of the nonmatched persons on Census Day that could be used later to determine their true residence statuses. Cognitive studies of the 1995 PES interview process revealed that many respondents misinterpreted the intent of the reconciliation process and reacted defensively to questions regarding discrepancies between the Census and PES rosters. Thus, the reconciliation procedure was changed in 1996 to an approach which was believed to be much less confrontational.

In the 1996 PES test was conducted in Chicago, Illinois, some months after Census Day also using CAPI methods. The rostering procedures were identical to 1995 except that immediately after collecting the household roster and prior to matching the PES and Census lists, the interviewer collected key residency data that would be needed for resolving roster discrepancies. This was done to avoid challenging the PES respondent's reports of who lived at the address on Census Day. During the reconciliation process, respondents were queried only about individuals listed on the Census roster who were not rostered in the PES. Another important difference between the two years was the initial rostering procedures for the PES. In 1995, PES respondents were asked only about individuals staying at the address on Census Day. In 1996, a modified procedure was used which listed all persons who stayed at the household the previous evening and used this information to help reconstruct the Census Day household.

In addition, in 1996 the interviewers were not able to jump back to the PES roster and make changes once the original Census roster was revealed as they were in 1995. This change may have resulted in greater independence between the Census and PES rostering processes and greater accuracy of the reconciled roster information. There were several other improvements implemented in the 1996 evaluation reinterview procedures based upon lessons learned in 1995 which are documented in Biemer, Woltman, Raglin, and Hill (1999).

2.2. The evaluation survey design

In both years, a reconciled reinterview survey was conducted following the PES to evaluate the accuracy of the PES residence status classifications. In 1995, the reinterview occurred almost immediately following the PES; however, in 1996 it was delayed for about five months following the PES. Since the reference point for the evaluation survey is Census Day, the delay in 1996 has the potential for increasing recall error in constructing a roster for Census Day household members. As we shall see in the presentation of

the results, this delay may be one of the primary reasons PES data quality appears worse in 1996 than in 1995 as reported by Raglin et al. (1998).

In both years, the survey was administered to a sample of PES respondents using a CAPI questionnaire that was essentially identical to the PES questionnaire. The evaluation reinterview proceeded in the same manner as the PES interview except that the roster obtained in the reinterview was matched against a roster of all individuals who were previously reported as living at the address by either the Census or the PES, referred to as the Census-PES combined roster. The reinterview enumerator reconciled all discrepancies between the newly obtained roster and the combined roster using procedures identical to those based in the PES.

The key differences between the PES and the evaluation survey data collection operations were in data collection and processing staff. The U.S. Census Bureau attempted to employ the most qualified and skilled enumerators and other operations staff available from the PES operation in the evaluation survey in order that the results of the reinterview could be treated as a gold standard for evaluating PES error.

In both 1995 and 1996, the evaluation survey sample was selected using a stratified cluster sample. PES housing units were first partitioned into six strata based upon the number of Census and PES roster nonmatches. Stratum 1 contained all households where all individuals rostered in the Census and the PES matched. Strata 2, 3, and 4 contained households where at least one person matched and one person on either the Census or the PES roster or both did not match. These strata will be combined in the analysis below. Stratum 5 contained whole household nonmatches with no individuals on the Census roster and Stratum 6 contained whole household nonmatches with at least one person on the Census roster. Since most rostering errors are expected to come from Strata 2, 3, and 4, these strata were oversampled to reduce the variance of an estimator based on the number of errors. Table 1 shows the final samples sizes by stratum for both years.

The sample sizes for the 1995 and 1996 evaluation reinterviews were 947 and 869 housing units, respectively. Ineligible cases such as duplicates, vacant units, nonhousing units, and units reinterviewed for quality control purposes were removed after the samples were selected. The total number of individuals rostered across all three systems was 2,963 for 1995 and 4,095 for 1996. The higher number in 1996 was primarily due to the rostering method used in the PES and the evaluation reinterview (described above) which asked

Table 1. Sample design

Sampling stratum	Analysis stratum	Final sample size	
		1995	1996
I	1	116	104
II	2	153	184
III		195	165
IV		83	67
V	3	238	174
VI	4	162	175
Total		947	869

respondents to name people staying at the address the previous evening and then use this list to construct the final roster.

The evaluation reinterview was designed to provide a picture of the accuracy of the PES data collection and processing operations. It was used by the Census Bureau to examine various components in the DSE. Analyses of PES enumeration bias assuming that the evaluation interview provides a gold standard measurement is given in West and Griffiths (1996) for the 1995 evaluation and Raglin et al. (1998) for the 1996 evaluation. In what follows, we reanalyze the evaluation data for these two tests using LCA. The advantage of this approach is that, unlike the analysis conducted by West and Griffiths (1996) and Raglin, Griffin, and Kromar (1998), it is not necessary to assume a gold standard. Using LCA, the error rates associated with all three classification systems – Census, PES, and the evaluation reinterview – can be estimated. It is important to note that, although closely related to triple system estimation models found in the capture-recapture literature, there are some important differences between our models and objectives and those used in census population size estimation. These differences will be discussed at the end of the next section.

3. The Technical Approach for PES Evaluation

In this section, we develop the statistical framework for modeling the enumeration error in the Census, the PES, and the PES evaluation reinterview survey. Each of these data collection systems starts with a roster of all individuals who are potential members of a household. For the Census, the process ends once the household members are listed on the Census form. For the PES and evaluation reinterview, the process begins with an initial listing of potential household members, but continues until each person is verified as either a Census Day resident, Census Day nonresident, or unresolved, meaning there was insufficient information collected to determine a residence status and, thus, the status must be imputed. Thus, the universe for the three systems is the union of three overlapping lists of individuals: individuals listed on the Census roster and individuals listed on the PES or evaluation survey rosters prior to the reconciliation process.

Therefore, let H denote the set of all households in the area to be enumerated and let h denote a particular household in H. Let U_h denote the individuals in household h who would be listed as residents of the household in either the Census, the PES, or the evaluation survey. This includes actual residents as well as nonresidents and various types of erroneous enumerations. Let $\wp_h \subseteq U_h$ denote all persons who are truly residents of household of h. Define $U = \bigcup_h U_h$ as the universe of potential population members and $\wp = \bigcup_h \wp_h$ as the subset of U composed of individuals who are true residents of the households in H and should be counted; i.e., U denotes the population of all individuals who would be named as living in households in a test census site and $\wp$ is the set of individuals who are truly residents of the households in the site. The component of $\wp$, (i.e., $U \sim \wp$) is the set composed of persons who are not residents of the households in H and other erroneous enumerations.

Let (h, i) denote the ith person identified in household h and let $X_{hi} = 1$ if $(h, i) \in \wp_h$ and 2 otherwise. Thus, X_{hi} is an indicator variable defined for a person in U_{hi} that denotes membership in $\wp_h$. We assume that X_{hi} is latent (unobserved) variable and that the Census,

PES, and evaluation reinterview provide fallible indicators of X_{hi}. The goal of our analysis, then, will be to use these three indicators or residency classifiers to model the distribution X_{hi} as well as the error parameters associated with each classifier.

3.1. Definition of the classifiers

As described above, the Census classification process is conducted by a household respondent (in completing the census questionnaire) or by an enumerator working with the respondent (in the Nonresponse Followup interview). The respondent considers the potential member of the household and, using the census instructions as a guide, either lists the potential member of the roster, and thereby classifies the person as in $\wp$, or does not list the potential member, thus classifying him/her as not in $\wp$ by default.

We denote the Census classification variable as A_{hi} for $(h, i) \in U$. Thus, A_{hi} is a dichotomous indicator of true classification X_{hi} and is defined as $A_{hi} = 1$ if (h, i) is listed on the Census roster for household h and 2 otherwise.

The PES interviews are conducted in person by enumerators using CAPI. First, a new Census Day roster is obtained from the respondent by the enumerator without reference to or knowledge of the contents of the original Census roster. This new (''independent'') roster is then compared with the Census roster for the household and any differences are noted and are eventually discussed and reconciled with the respondent. The result of this reconciliation process is a classification of each person on the combined Census and PES roster as either a resident, a nonresident, or unresolved. Thus, two rosters are generated by the PES process. The first is the roster obtained prior to reconciliation and the second is the final, reconciled roster generated by the reconciliation process.

First, we define a classifier, P_{hi}, for the pre-reconciled PES roster. For $(h, i) \in U$, define $P_{hi} = 1$ if (h, i) is listed on the PES roster prior to reconciliation and 2 otherwise. Then, we define a classifier, B_{hi}, for the resolved PES roster (reconciled roster). For $(h, i) \in U$, define the PES reconciled classifier which has four states as follows:

$$B_{hi} = \begin{cases} 1 & \text{if classified as a resident} \\ 2 & \text{if classified as a nonresident} \\ 3 & \text{if classified as unresolved or status unknown} \\ 4 & \text{if not rostered} \end{cases}$$

The first three categories are self-explanatory. The fourth category applies to individuals who are not rostered by either the Census or PES process but are later identified in the evaluation reinterview. Since these individuals were included on the Census or the PES roster, we assigned a code of ''Not Rostered'' to them prior to analysis. There are a substantial number of these individuals in both tests censuses and we will be particularly interested in determining whether these individuals are people who should have been enumerated and classified as residents by the PES.

As mentioned previously, the evaluation reinterview is actually a second reinterview of the census respondents and uses procedures very similar to those of the PES interview. In the evaluation reinterview, a third roster of the Census Day residents is constructed by household respondents using free recall and without referring to the previous two rosters. The evaluation reinterview roster is compared to the combined Census and PES

roster and any differences are reconciled with the respondent. The result of this reconciliation process is the evaluation reinterview classification, C_{hi}, defined for $i \in U$ as:

$$C_{hi} = \begin{cases} 1 & \text{if classified as a resident} \\ 2 & \text{if classified as a nonresident} \\ 3 & \text{if classified as unresolved or status unknown} \end{cases}$$

Unresolved cases are handled in the same manner as described for the PES.

3.2. *Model assumptions and notation*

Let E_{hi} and F_{hi} denote any two arbitrary latent or manifest variables defined for $(h, i) \in U$, let $\boldsymbol{\pi}_{ef(h,i)}$ denote $\Pr(E_{hi} = e, F_{hi} = f)$. The conditional probability $\Pr(E_{hi} = e|F_{hi} = f)$ is denoted by $\boldsymbol{\pi}_{e|f(hi)}$. In what follows we will assume that

$$\boldsymbol{\pi}_{e|f(hi)} = \boldsymbol{\pi}_{e|f}$$

that is, the classification probabilities are homogeneous across individuals and households within the levels of the variable or set of variables represented by F.[4] Thus, for notational convenience we will drop the subscript (h, i) when it is clear we are referring to an individual in the universe.

Let *XAPBC* denote the cross-classification table for the variables *X,A,P,B,* and *C* for all $(h, i) \in U$ and let (x, a, p, b, c) denote the cell associated with $X = x$, $A = a$, $P = p$, $B = b$, and $C = c$ in this table. Define $\boldsymbol{\pi}_{xapbc}$ as the expected proportion in cell (x, a, p, b, c). Then, we can write $\boldsymbol{\pi}_{xapbc}$ as

$$\begin{aligned} \boldsymbol{\pi}_{xapbc} &= \Pr(X = x) \Pr(A = a|X = x) \Pr(P = p|A = a, X = x) \\ &\quad \times \Pr(B = b|A = a, P = p, X = x) \Pr(C = c|A = a, P = p, B = b, X = x) \\ &= \boldsymbol{\pi}_x \boldsymbol{\pi}_{a|x} \boldsymbol{\pi}_{p|ax} \boldsymbol{\pi}_{b|a\ px} \boldsymbol{\pi}_{c|apbx} \end{aligned} \tag{1}$$

The above identity demonstrates that the probability that an individual in U is classified in cell (x, a, p, b, c) can be decomposed into the product of marginal and conditional probabilities. Since the true classification, X, for the individual is assumed to be unobservable, it will be treated in the subsequent analysis as a latent variable. Thus, the full table *XAPBC* is also unobservable and, using the estimation methods developed for missing data problems such as the EM algorithm (Dempster, Laird, and Rubin 1977), *XAPBC* can be estimated under an assumed model for (1).

Note if we were to try to estimate all 95 parameters of the model in (1) we would need a minimum of 95 degrees of freedom. Since only 47 degrees of freedom are available, such a model would be unidentifiable.[5] To overcome this problem, restrictions on the probabilities must be introduced to reduce the number of parameters associated with the model.

[4] This assumption is similar to the assumption made for the DSE (see Wolter 1986). However, as we shall see, correlations between the errors in the classifiers *A*, *B*, and *C* are estimable in our analysis since *F* may be either a grouping variable or another classifier. Thus, the homogeneity assumption is less restrictive in our analysis.

[5] The rules for determining the number of parameters in an ANOVA model can also be used to determine the number of parameters in a latent class model (see, for example, Kempethorne 1975). In (1), the numbers of parameters associated with the five conditional probabilities on the right-hand side are (from left to right): 1, 2, 4, 24, and 68 with sum 95. The degrees of freedom for estimation is the number of cells in the *APBC* table minus 1 or $2 \times 2 \times 4 \times 3$ or 48.

The traditional latent class model for the table *XAPBC* introduces the restrictions which specify that the classifiers *A*, *P*, *B*, and *C* are mutually independent given the true classification *X*, or mathematically, that

$$\boldsymbol{\pi}_{p|ax} = \boldsymbol{\pi}_{p|x}, \boldsymbol{\pi}_{b|apx} = \boldsymbol{\pi}_{b|x} \quad \text{and} \quad \boldsymbol{\pi}_{c|apbx} = \boldsymbol{\pi}_{c|x} \tag{2}$$

and, thus, substituting these restrictions in (1) produces the probability model

$$\boldsymbol{\pi}_{xapbc} = \boldsymbol{\pi}_x \boldsymbol{\pi}_{a|x} \boldsymbol{\pi}_{p|x} \boldsymbol{\pi}_{b|x} \boldsymbol{\pi}_{c|x} \tag{3}$$

which contains 15 parameters and all parameters are estimable from the observed *APBC* table.

Note that the term $\boldsymbol{\pi}_{a|x}$ may be interpreted as the error probability for indicator *A* since the probabilities $\Pr(A = 1|X = 2)$, the false positive error probability, and $\Pr(A = 2|X = 1)$, the false negative probability, are both represented in $\boldsymbol{\pi}_{a|x}$. Likewise, $\boldsymbol{\pi}_{p|x}$, $\boldsymbol{\pi}_{b|x}$, and $\boldsymbol{\pi}_{c|x}$ are error terms for the classifiers, *P*, *B*, and *C*, respectively. Thus, the restrictions in (2) represent the independent classification errors model.

The software that will be used for the analysis requires that latent class models be specified as log-linear models. Haberman (1979) demonstrated that (3) is equivalent to the hierarchical log-linear model with highest order terms XA, XP, XB, and XC. This model is sometimes written in shorthand notation as $\{\mathrm{XA}, \mathrm{XP}, \mathrm{XB}, \mathrm{XC}\}$. All the models considered in the subsequent analysis will be hierarchical models; i.e., models for which all lower terms that involve variables in the higher order interaction terms are included.

Some of the models that we will explore can only be expressed as ''modified path models'' (Goodman 1974). These are essentially models formed by replacing each of the conditional probabilities terms in (1) by a logistic regression model for the probability. One advantage of this method is that, unlike log-linear models, the causal ordering of the variables can be explicitly represented in the path model formulation of the probability model. For the present application, the Census classification occurs first chronologically followed by the PES unreconciled classification, the PES reconciled classification, and finally, the evaluation reinterview classification. Therefore, a modified path model for the probability expression in (3) replaces the conditional probabilities for A|X by the logistic model {AX}; P|X by {PX}; B|X by {BX}; and C|X by {CX}. More complex expressions can be developed from (1) by replacing each of the conditional probabilities with a corresponding logistic model with restrictions on the terms to obtain identifiability.

Finally, we consider the addition of the stratification variable to the models. Recall that in selecting the sample for the evaluation reinterview, six strata were formed based upon agreement between Census classification, *A*, and the PES, *B*. The evaluation reinterview sample was then selected by simple random sampling within these strata. We introduce the stratification variable, *S*, defined in Section 2 into the model to account for any variation in the enumeration probabilities by stratum. Define the variable *S* as

$$S_i = \begin{cases} 1 & \text{if observation i} \in \text{stratum I} \\ 2 & \text{if observation i} \in \text{stratum II, III, or IV} \\ 3 & \text{if observation i} \in \text{stratum V} \\ 4 & \text{if observation i} \in \text{stratum VI} \end{cases}$$

Then, introducing S into identity in (1), we write

$$\pi_{sxapbc} = \pi_x \pi_{a|x} \pi_{p|xa} \pi_{b|xap} \pi_{s|xabp} \pi_{c|xapbs} \tag{4}$$

Note that, since A, P, and B precede S causally (i.e., S is formed after A, P and B are observed), then A, P, and B given X are assumed not to depend on S. However, C depends upon S since the partitioning of the sample by S precedes the determination of C by design. Further, since the strata were formed without regard to the prereconciled classification, P, we will assume $\pi_{s|xabp} = \pi_{s|xab}$; that is, S is independent of P given X, A, and B. Further restrictions on (4) are necessary since the number of parameters implied by (4) exceeds the number of degrees of freedom for estimation.

3.3. Correspondence with dual system and triple system estimation

There is a considerable literature which addresses the estimation of the total size of the Census population when data on Census Day residency is available from two or more sources. The dual system estimation methodology currently used by the U.S. Census Bureau is documented in Wolter (1986). However, it is well-known that this estimator may be considerably biased if enumeration or ''capture-recapture'' probabilities vary within the domains defined by the model or if the enumeration errors for the two systems are correlated. This bias is known as ''correlation bias'' (see, for example, Wolter 1986; Mulry and Spencer 1993). Several researchers have developed methods for estimating population size when three enumeration systems are available such as the Census, the PES, and an administrative records data base or population registry (for example, Zaslavsky and Wolfgang 1993; Darroch, Fienberg, Glonek, and Junker 1993). With triple system estimation, the correlation bias can be estimated and accounted for in estimates of population size under some assumed models.

The methodology developed in this article also uses three systems (viz., the Census, the PES, and the evaluation reinterview survey) and is closely related to the dual and triple system estimation methods, but there are some important differences. For example, the literature on capture-recapture census population estimation assumes that people in the target population are either counted (''in'') or missed (''out'') by the enumeration system. Then a 2^3 cross-classification table can be formed that classifies individuals as either ''in'' or ''out'' in all three systems. One of the cells is necessarily unobserved in the table corresponding to individuals who are missed in all three systems.

In our formulation, we are not primarily interested in estimating total population size. Rather, we are interested in how accurately each system classifies individuals who are rostered as either Census Day residents, nonresidents, or unresolved. Note that in our formulation there is no empty cell corresponding to individuals missed in all three systems since we confine our analysis to U, i.e., the universe of people who are rostered by the systems in the analysis. This is because our models operate only for individuals who are listed on at least one roster in the Census, PES, and evaluation reinterview rostering processes. Triple system estimation for census population size estimation projects an estimate for individuals who are Census Day residents who are not listed in the rostering process as well as individuals who are listed. Thus, the target populations for the two approaches are quite different.

The methods we employ are straightforward applications of log-linear models with latent variables (see Hagenaars 1993 for an introduction to these models). However, special methods must be employed for triple system population size estimation due to the presence of the structural 0 for the cell corresponding to being missed in all three systems.

Another distinction between our models for estimating enumeration error and the capture-recapture models is the assumption of a two-class latent variable for resident and nonresident. The capture-recapture models in the literature essentially assume a single-class latent variable corresponding to Census Day resident. That is, terms for erroneous enumerations are not incorporated into the models, but rather it is assumed that all erroneous enumerations have been removed from all three systems prior to estimation. Work is now underway at the U.S. Census Bureau to apply latent class models for multiple system census population estimation using models similar to those developed here which take into account the potential for erroneous enumerations in all three systems (Biemer 2000).

4. Analysis of 1995 and 1996 PES Evaluation Data

As described in Section 2, the evaluation reinterview sample is a stratified random subsample of the PES sample with six strata. Because of the similarity of sampling strata II, III, and IV, the sample units in these strata were combined into a single stratum in the analysis. This modification also reduced the effect of sparse cells on the fit statistics for the analysis. Prior to analysis, the data for both years were appropriately weighted for the evaluation reinterview probabilities of selection and the weights rescaled so that they totaled to the evaluation reinterview sample size. Latent class models were fitted to the weighted data using the ℓEM Version 1.0 software package (Vermunt 1997). All models were run multiple times with different starting values to verify identifiability and to check for local minima.

In the following analysis, the variables A, P, B, and C denote the Census, prereconciled PES, final reconciled PES, and final reconciled reinterview classification, respectively; X denotes the latent true classification, and S denotes the stratum indicator. Note that, by design, $\Pr(B = 4|A = a, P = b) = 0$ for $(a, b) = (1, 1)$, $(1, 2)$, or $(2, 1)$. That is, persons who appear on the Census roster or the PES independent roster cannot receive a ''Not Rostered'' code in the PES reconciled roster since these persons were carried forward to the PES reconciled roster. Thus, these additional constraints were added to the specifications of all the models considered.

Surprisingly, for both 1995 and 1996, models that allow for variation across sampling strata in the classification errors were rejected in favor of models that specify homogeneous classification errors within strata. This suggests that the sampling strata used for drawing the reinterview survey sample were not very effective at reducing the variation in classification probabilities for the evaluation reinterview. Therefore, all the models we consider in the next section omit any interaction terms involving the C by S interaction.

4.1. Model selection for 1995 and 1996 PES analysis

Table 2 summarizes the model fit statistics for four basic models that were fit to data for both years data. Model 1 is the simple latent class model that assumes independent

Table 2. Model selection results for 1995 PES evaluation data

Model	Effects	df	npar	1995 Test census			1996 Test census		
				L^2	p	BIC	L^2	p	BIC
1	A|X = {AX} P|XA = {PX} B|XAP = {BX} S|XAB = {SX SA SB} C|XAPBS = {CX}	158	33	263	0.00	–1086	1017	0.00	–298
2	Model 1 except B|XAP = {BX, BP} C|XAPBS = {CX, CB}	148	43	81	1.00	–1102	209	0.00	–1022
3	Model 2 except B|XAP = {BX, BP, BA}	145	46	77	1.00	–1082	93	1.00	–1113
4	Model 3 except C|XAPBS = {CX, CB, CA}	143	48	75	1.00	–1068	85	1.00	–1104
	Model 1–Model 2	10	10	182	0.00	16	808	0.00	724
	Model 2–Model 3	3	3	4	0.26	−20	116	0.00	91
	Model 3–Model 4	2	2	2	0.37	−14	8	0.02	−9

Note 1: The symbol V|X . . . Y before the submodel specification denotes the conditional probability in (1).
Note 2: The best model is the model having the smallest BIC and $p > .05$.

classification errors for all four classifiers. (i.e., no interactions among the error terms AX, PX, BX, and CX) and homogeneity of error rates across strata. Model 3 is referred to as a first order dependence model since it specifies an interaction (or correlation) between each classifier and the classifier that just precedes it. Thus, Model 2 differs from Model 1 by the addition of the terms PA, BP, and CB. Model 2 is an extension of the dependence model to include additional between two consecutive classifiers in the Census/PES/reinterview process. Finally, Model 4 further extends the independent classification error model to include interactions between the reinterview classifier (C) and the Census classifier (A).

The criteria we used for identifying the best model are: fit, parsimony, and plausibility of the estimates. To determine the fit of a model to the observed data, we used the log-likelihood goodness of fit criterion. The higher the chi-squared p-value, the better the fit of the model. A typical rule of thumb used in the log-linear modeling literature is that the p-value for the model L^2 should exceed 0.05 for an acceptable model fit where L^2 is the value of the likelihood ratio chi-squared statistic. However, there may be a number of models that satisfy this criterion and the ideal model is the most parsimonious model. Therefore, a second criterion was the Bayesian Information Criterion (BIC) defined as $L^2 - (\log N)\, df$ where N is the sample size, and df is the degrees of freedom for the fitted model. Among all models that fit the data (i.e., p-value > 0.05) the best model is the one whose BIC is minimized (see Liu and Dayton 1997 for a justification of this approach).

The third criterion, plausibility of a model, is a much more subjective criterion. First, only models that were consistent with the sampling design and data collection methodology used in the test censuses were considered in our analysis. For example, we did not consider models for the component Pr(A|X) that include terms for B or C since these

measurements were collected later in the PES-evaluation reinterview process. In addition, models that gave reasonable and consistent estimates of classification error rates were preferred over models whose estimates were highly unlikely or gave results which were inconsistent with other external information on the error rates.

Applying these criteria to the models for 1995, we see that Models 2, 3 and 4 fit the data very well as in all three cases the log-likelihood chi-squared statistic has a p-value of approximately 1.00. However, the BIC measure suggests that Model 2 is best and, as we will see, Model 2 yields plausible estimates of the parameters. Further, the simple likelihood ratio tests comparing the models indicated no significant improvement in fit beyond Model 2. Thus, we will use Model 2 to generate the estimates of the classification probabilities for the Census, PES, and evaluation reinterview.

From the model comparisons in the bottom rows in the table, note that the terms BA and CA are not significantly different from 0. This follows since the differences between Models 2 and 3 and Models 3 and 4 are not significant. This implies that neither PES errors nor errors in the evaluation reinterview are correlated with the Census errors. Since BA is somewhat related to the correlation bias term in the DSE, the fact that BA is not statistically significant from 0 indicates that classification errors in the Census and PES are not correlated for persons in U. Thus, this type of correlated error would appear not to be an important factor in the DSE correlation bias. However, the term CB in Model 2 is highly significant, indicating that the evaluation reinterview errors are correlated with the PES enumeration errors. This may suggest that the evaluation reinterview interviews are being influenced by the PES classification of rostered individuals.

For the 1996 data set, Model 2 does not fit the data well and the choice is between Models 3 and 4. This suggests that, in 1996, there was a significant correlation between the errors in PES and those in the Census errors, violating the assumption of the DSE. Using the BIC criterion, Model 3 provides the most parsimonious fit; however, the term CA in Model 4 is statistically significant ($\alpha = 0.05$). Further analysis indicates that the classification probability estimates for Models 3 and 4 do not differ appreciably so the choice is somewhat arbitrary. For the sake of consistency with our selection criteria, we will use Model 3 to estimate the classification error probabilities for 1996.

4.2. Estimates of PES error probabilities

Classification error probabilities were estimated for the 1995 PES under Model 2 and for the 1996 PES under Model 3 using the ℓEM software. The maximum likelihood estimates appear in Table 3 as well as the estimates from the 1995 and 1996 traditional analysis (see West et al. 1996; Raglin et al. 1998). Standard errors for the LCA estimates were approximated by the corresponding standard errors from the traditional analysis so caution should be exercised when interpreting the comparisons in the table.

Table 3 is split into two halves corresponding to the 1995 test and the 1996 test, respectively. For each year, the true classification is shown across the columns and the observed PES classification is shown down the rows. Within each cell of the true status by PES status classification, we report both the estimates from the LCA and those from the traditional analysis.

It is obvious from the table that LCA and traditional estimates are quite different

Table 3. Comparison of LCA and traditional estimates of PES classification error rates for 1995 and 1996 Census tests (standard errors in parentheses)

PES Classification *n*	1995 PES data				1996 PES data			
	True classification				True classification			
	Resident		Nonresident		Resident		Nonresident	
	LCA	Traditional	LCA	Traditional	LCA	Traditional	LCA	Traditional
Resident	98.2 (0.8)	90.0 (0.8)	59.4 (4.3)	28.8 (4.0)	95.8 (2.2)	79.7 (2.2)	41.0 (3.4)	58.5 (3.4)
Non resident	0.0 (0.4)	1.7 (0.4)	8.9 (0.6)	23.5 (3.7)	3.7 (0.5)	4.0 (0.5)	22.1 (3.7)	34.8 (4.2)
Unresolved	1.8 (0.5)	3.2 (0.5)	8.3 (1.7)	12.1 (2.1)	0.5 (0.4)	1.3 (0.4)	3.1 (0.4)	1.5 (0.4)
Not rostered	0.0 (0.5)	5.1 (0.5)	23.4 (4.2)	35.6 (4.3)	0.0 (2.1)	15.1 (2.1)	33.7 (4.2)	5.2 (0.5)

Notes: Entries are the conditional probabilities of the PES classification (rows) given the true classification (columns). The missing error rate is the sum of two cells: true resident/PES nonresident and true resident/PES not rostered. The false positive rate is the cell corresponding to true nonresident/PES resident.

and lead to very different conclusions regarding the overall quality as well as the relative quality of the 1995 and 1996 PES data. For example, in the 1995 PES, the estimated proportion of true residents who were either classified as ''nonresident'' or ''not rostered'' is 0.0 for LCA and 6.8 percent for the traditional analysis (computed as the sum of ''nonresident'' and ''not rostered'' categories). Further, the traditional analysis suggests that a substantial proportion (5.1 percent with a s.e. is 0.5 percent) of the true residents were not rostered by the PES; however, the LCA estimate of this proportion is 0, indicating that missing individuals in the PES was not a problem.

Note that an estimate of 0 for the probability of classifying a true resident as ''nonresident'' or ''not rostered'' does not indicate that the PES does not miss any true Census Day residents. Rather, this LCA result suggests that the PES did not miss any true residents who were identified by either the Census or the evaluation reinterview, or conversely, that the evaluation reinterview did not identify any new residents that were not already identified by either the Census or the PES. This interpretation follows directly from our definition of the universe for the study, U. Our analysis, therefore, does not provide an estimate of the total number of residents missed by the PES.

The 1995 traditional and LCA results for true nonresidents in the population are also inconsistent. From Table 3 we note that the probability a true nonresident is classified as a ''resident'' (i.e., a false positive error[6]) is 59.4 percent (s.e. of 4.3 percent) for LCA but only 28.8 percent (s.e. of 4.0 percent) for traditional analysis. Thus, the LCA analysis indicates that false positive errors are much more frequent in the PES than the traditional analysis indicates they are.

[6] The false positive rate is closely related to the ''erroneous enumeration rate'' often cited in the decennial census evaluation literature (see, for example, Mulry and Spencer 1993). The difference is that, while the denominator of the false positive rate for a classifier is the number of true nonresidents in U, the denominator of the erroneous enumeration rate is the estimated number of persons in $\wp$ based upon the classifier.

There are important differences between LCA and traditional analysis for 1996 as well. In Table 3, we see from that the LCA estimate of the proportion of true residents who were classified as either ''nonresident'' or ''not rostered'' in 1996 is only 3.7 percent for LCA compared with 19.1 percent for the traditional analysis. (As before, these estimates are computed as the sum of the ''nonresident'' and ''not rostered'' categories). Further, traditional analysis indicates that a very large proportion (15.1 percent with s.e. of 2.1 percent) of the true residents were not rostered by the PES which suggests an even worse problem of missing residents in the 1996 PES than was observed for the 1995 PES. By contrast, the LCA estimate of this proportion is 0 which is the same as the 1995 PES LCA estimate.[7]

For the true nonresidents in the population in 1996 LCA estimate of the false positive rate is lower than the rate from traditional analysis: 41.0 compared with 58.5. Also, LCA suggests that a large percentage (33.7) of the nonresidents are actually individuals who were not rostered in the PES. Further analysis revealed that these are individuals who were listed for the first time in the evaluation reinterview, which suggest that traditional analysis and LCA are very different in their treatment of individuals who were identified only in the evaluation reinterview.

4.3. Comparisons between 1995 and 1996

In considering the differences between the 1995 and 1996 PES error rates in Table 3, it is important to recall that the target populations for the two years were very different. In 1996, difficult to enumerate households were oversampled in the Chicago area whereas this was not the case in 1995 for the Oakland sample. Thus, the task of determining true Census Day residency was expected to be much more difficult in 1996 than in 1995.

One indicator of the relative enumeration difficulty of the two sites is the Census correct enumeration rate; i.e., the proportion of individuals identified on the Census roster who are true Census Day residents. From the LCA, we estimated the proportion of individuals identified in all three systems that were true residents of the households in which they were found to be 69.3 percent in 1995 and 61.9 percent in 1996 (significant at $\alpha = 0.10$). Thus, a significantly larger percentage of individuals identified in the Census process were nonresidents in 1996, as expected.

Table 4 summarizes some other results from the LCA related to the accuracy of the Census enumeration. As seen from Table 4, the Census miss rate was significantly higher in 1996 than in 1995: 23.9 percent (s.e. of 2.0 percent) compared with 10.0 percent (s.e. of 4.0 percent). This is also an indication that the 1996 site contained a higher proportion of hard to enumerate individuals than the 1995 site.

Thus, in comparing the two years, we would expect error rates to be somewhat better in 1995 than in 1996 even if the quality of the PES process was the same for both years. For this reason, any improvement in the 1996 PES suggested by the estimates in Table 3 probably understates the true improvement in the PES process. By the same reasoning, any deterioration in quality suggested by these data may also be somewhat overstated.

Note that the estimates in Table 3 from traditional analysis suggest that PES data quality was significantly worse in 1996 than in 1995. For example, the PES miss rate (i.e., the

[7] Raglin et al. (1998) also suggested that the extra people found by the 1996 PES evaluation reinterview and classified as residents were unlikely to be residents.

Table 4. LCA estimates of Census classification error for the 1995 and 1996 test evaluations (standard errors in parentheses)

Census classification	1995 data		1996 data	
	True classification		True classification	
	Resident	Nonresident	Resident	Nonresident
Rostered	90.0 (4.0)	15.6 (2.3)	76.1 (2.0)	28.8 (1.8)
Not rostered	10.0 (4.0)	84.4 (2.3)	23.9 (2.0)	71.2 (1.8)

Notes: Entries are conditional probabilities of the Census Classification (rows) given the true classification (columns). The missing error rate is the cell true resident/Census not rostered. The false positive rate is the cell corresponding to true nonresident/Census rostered.

probability of classifying a true resident as either ''nonresident'' or ''not rostered'') increased from 6.8 percent (i.e., the sum of 1.7 and 5.1) in 1995 to 19.1 percent (i.e., the sum of 4.0 and 15.1) in 1996, according to the traditional analysis. In addition the false positive rate (i.e., the probability of assigning a resident code to a true nonresident) increased from 28.8 percent to 58.5 percent. These differences are both highly statistically significant. LCA, however, provides a very different picture of the differences between the two years. In fact, the LCA method suggests that there is actually no significant difference in the accuracy with which the PES classified persons who are true Census Day residents, which by our foregoing discussion regarding the relative difficulty of the two evaluation sites would suggest improvement in 1996. For true nonresidents, there is further improvement according to the LCA estimations, with only 41.0 percent false positives in 1996 compared with 59.4 percent false positives in 1995 (significant at the $\alpha = .05$ level).

It is difficult to reconcile the differences between the traditional analysis and LCA. Which set of results is more indicative of the true PES process quality? There are a number of reasons to doubt the results of the traditional analysis. First, consider the proportion of census residents that the traditional analysis indicates were completely missed, i.e., not rostered, by the PES but were rostered by the evaluation reinterview survey. According to the traditional analysis, this number was approximately 5.1 percent of the rostered residential population in 1995 and 15.1 percent in 1996 (i.e., in Table 3 under the Resident-Traditional column, the ''not rostered'' percentage went from 5.1 percent in 1995 to 15.1 percent in 1996). The 1996 result seems implausible since the evaluation reinterview applied the same methods for rostering as the PES applied. Therefore, it is difficult to understand how these rostering procedures failed to identify a large number of individuals in the PES but were successful at identifying the same individuals in the evaluation reinterview, especially since the evaluation reinterview was conducted five months after the PES and, thus, five months further from Census Day.

One possible explanation is that the evaluation reinterview misclassified these new individuals as residents. This could occur, for example, if the evaluation reinterview identified a number of individuals who moved in after the PES was conducted and failed to properly determine their in-mover statuses. The fact that the proportion of not rostered individuals in the PES tripled from 1995 to 1996 is consistent with the ''in-mover theory''

since the time lag between the PES and the evaluation reinterview was only one month in 1995 compared with five months in 1996. Thus, it is possible that the proportion of movers in the target population was much greater for the 1996 Census Test. In addition, the 1996 test evaluation site (Chicago) contains several major universities and with Census Day in occurring in October and the evaluation reinterview occurring the following summer, college students would provide an abundant source of movers in the 1996 reinterview survey. Anecdotal evidence from the data collection operations also supports this explanation for the increase in PES ''not rostered'' individuals (see Raglin et al. 1998).

In the next section, we consider the accuracy of the evaluation reinterview and provide evidence from LCA that the reinterview process was itself subject to large classification errors.

4.4. The accuracy of the evaluation reinterview results

As a by-product of the latent class analysis of the PES process, estimates of evaluation reinterview classification error are also generated. Table 5 provides the classification error estimates for both years. Note that the miss rate is significantly larger for the 1996 reinterview, according to LCA – 13.3 percent compared with 0.0 percent in 1995. Further, the probability of correctly classifying nonresidents is significantly higher in 1996 than in 1995. A process that always classifies individuals as residents would have a zero miss rate and a high false positive rate. Likewise, a process that always classifies individuals as nonresidents would make no false positive errors but would have a large miss rate. Thus, the results in Table 5 suggest a tendency in the 1996 evaluation reinterview to classify individuals as nonresidents and a tendency in the 1995 evaluation reinterview to classify individuals as residents.

Another factor in the comparison of false positive rates in 1995 and 1996 is the difference in the nonresidential populations identified by the two tests. Recall that both the 1996 evaluation procedures were designed to include a larger number of potential Census Day residents than in 1995. As discussed previously, the 1996 initial rostering

Table 5. Estimates of evaluation reinterview classification error for the 1995 and 1996 test evaluations (standard errors in parentheses)

Evaluation Reinterview Classification	1995 data		1996 data	
	True classification		True classification	
	Resident	Nonresident	Resident	Nonresident
Resident	97.2 (0.4)	76.9 (2.3)	83.7 (0.8)	62.2 (1.2)
Nonresident	0.0 (n/a)	9.9 (1.5)	13.3 (0.7)	33.4 (1.2)
Unresolved	2.8 (0.4)	13.2 (1.7)	3.0 (0.4)	4.4 (0.5)

Notes: Entries are the conditional probabilities of the reinterview classification (rows) given the true classification (columns). The ''missing'' or false negative error rate is the cell corresponding to true resident/reinterview nonresident. The false positive rate is the cell corresponding to true nonresident/reinterview resident.

procedure listed all individuals who stayed at the household the previous evening and used this information to help reconstruct the Census Day household. This would result in a somewhat higher proportion of nonresidents in 1996. This is also borne out by the LCA estimates of the proportion of the individuals in the target population who are nonresidents. For 1995 it was 30.7 percent (s.e. of 3.3) and in 1996 it was 39.1 percent (s.e. of 1.7).[8] In addition, many individuals rostered in the 1996 PES and PES evaluation reinterview may have much looser connections to the households than those rostered in 1995 and, as a result, can be more accurately classified as nonresidents. This would explain the higher accuracy rate for identifying nonresidents in both the PES and the reinterview in 1996.

5. Discussion

This article demonstrates the use of latent class analysis methods to evaluate the accuracy of census enumeration processes when no gold standard measurements exist. When the gold standard measurements are, themselves, subject to large enumeration errors, the errors in the Census and the PES can be considerably overstated. Latent class analysis offers an alternative method of analysis that may give very different estimates of the magnitudes of the errors as well as insights into the causes and origins of census error.

In our application of the methodology using the 1995 and 1996 Test Censuses, the estimates derived from LCA differed importantly from those obtained by the traditional approach. Using the traditional approach, the PES appeared to have much higher miss rates and the probability of a miss increased substantially in 1996 as compared to 1995. The source of this large increase in missed individuals was primarily the increase in people identified in the evaluation reinterview who where somehow missed by both the 1996 Census and PES.

However, LCA estimates provide a very contradictory picture of data quality in 1996, particularly for true residents. Not only were the LCA estimates of the PES correct enumeration probabilities considerably higher in both years than those from traditional analysis, but the differences between the estimates for the two years were smaller for LCA. Perhaps the biggest contradiction between the two sets of estimates occurs for the 1996 Test Census for the individuals who were rostered in the evaluation reinterview but not rostered in either the Census or the PES. While traditional analysis suggests that most of these individuals are true residents that were missed by the PES, LCA clearly indicates that these individuals are true nonresidents and were, therefore, correctly excluded in the PES.

With regard to the classification of nonresidents, the two methods also differed dramatically. For 1995, the traditional approach indicates a relative low rate of false positives which worsened considerably in 1996. However, the LCA estimates indicate that the false positive rate was significantly higher in both years and significantly improved in 1996.

The results of traditional analysis of the 1995 and 1996 Test Census would lead one to conclude that the improvements made in census enumeration methodology following the 1995 PES actually increased error rates in the 1996 PES. This is highly implausible

[8] Note: These estimates do not appear in the tables.

given that the 1995 research showed that the improvements would reduce PES error rates. The analysis using LCA is consistent with the direction of the improvements and are therefore more plausible.

These analyses provided several important lessons for the design in Census 2000. The key finding in this work is the considerable risk associated with using the traditional reconciled reinterview methodology for evaluating the quality of the PES. Our reanalysis of the 1995 and 1996 Test Censuses provides evidence of the fallibility of the evaluation reinterview process, a finding that is consistent with other research (see, for example, Biemer and Forsman 1992). In fact, the assumption that the evaluation reinterview process yields the true classification can be tested directly by constraining the reinterview false positive and false negative error rates to be 0 in the model fitting process. This test clearly indicates that the gold standard model does not fit the data on the basis of the chi-squared goodness of fit criterion and must be rejected. Our analysis provides strong evidence that evaluation reinterview data are subject to considerable classification error and are not suitable for use as a gold standard for evaluating the PES. However, LCA provides a means for assessing the quality of the PES data through the use of models that fit the observed data very well. Our results indicate that LCA is an important coverage evaluation methodology for census evaluations.

6. References

Biemer, P.P. and Forsman, G. (1992). On the Quality of Reinterview Data with Application to the Current Population Survey. Journal of the American Statistical Association, 87, 915–923.

Biemer, P.P. and Hubbard, M. (1996). Reconciliation Bias in Reinterview: Results from the SIPP Behavior Coding Study. RTI Final Report, Project 52U-5379-003.

Biemer, P.P., Woltman, H., Raglin, D., and Hill, J. (1996). The 1995 and 1996 Integrated Coverage Measurement Processes: An Evaluation Using Latent Class Analysis. U.S. Census Bureau Technical Report for Project 52U-5379-005, Design of the 1995 Test Census Evaluation, Washington, DC.

Biemer, P.P. (2000). Triple System Estimation with Erroneous Enumerations. U.S. Census Bureau Technical Report for Project 52U-6972-02, Latent Class Analysis Project, Washington, DC.

Brown, J., Diamond, I., Chambers, R., and Buckner, L. (1999). The Role of Dual System Estimation in the 2001 Census Coverage Surveys of the UK. Proceedings of Statistics Canada Symposium 1999, Ottawa, Canada.

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

Choi, C.V., Steel, D.G., and Skinner, T.J. (1988). Adjusting the 1986 Australian Census Count for Under-Enumeration. Survey Methodology, 14, 173–190.

Darroch, J., Fienberg, S., Glonek, G., and Junker, B. (1993). A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability. Journal of the American Statistical Association, 88, 1137–1148.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from

Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, B39, 1–38.

Goodman, L. (1974). Analysis of Systems of Qualitative Variables When Some of the Variables are Unobservable. Part I: A Modified Latent Structure Approach. American Journal of Sociology, 79, 1179–1259.

Haberman, S. (1979). Analysis of Qualitative Data: New Developments. New York: Sage Publications, Academic Press.

Hagenaars, J. (1993). Loglinear Models with Latent Variables. Newbury Park, CA: Sage Publications.

Hogan, H. (1993). The 1990 Post-Enumeration Survey: Operations and Results. Journal of the American Statistical Association, 88, 423, 1047–1071.

Kempethorne, O. (1975). The Design and Analysis of Experiments. Huntington, NY: Krieger Publishing Co.

Krosnick, J.A. and Alwin, D.F. (1987). Satisficing: A Strategy for Dealing with the Demands of Survey Questions. Paper presented at the annual meetings of the American Association for Public Opinion Research.

Kuha, J. and Skinner, C. (1997). Categorical Data Analysis and Misclassification. In Survey Measurement and Process Quality, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, New York: John Wiley and Sons.

Liu, T.H. and Dayton, C.M. (1997). Model Selection Information Criteria for Non-Nested Latent Class Models. Journal of Educational and Behavioral Statistics, 22, 249–264.

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

Raglin, D., Griffin, D., and Kromar, R. (1998). Internal Census Bureau Report on the 1996 Census Test Evaluations.

Sinclair, M. and Gastwirth, J. (1993). Evaluating Reinterview Survey Methods for Measuring Response Errors. Proceedings of the 1993 Annual Research Conference of the U.S. Census Bureau, Washington, DC, 771–738.

U.S. Bureau of the Census (1985). Evaluation of Censuses of Population and Housing. STD-ISP-TR-5, Washington, DC: U.S. Government Printing Office.

Vermunt, J. (1997). IEM: A General Program for the Analysis of Categorical Data. Tilburg University.

West, K. and Griffiths, R. (1996). Results from the 1995 Integrated Coverage Measurement Evaluation Interview. Proceedings of the 1996 Joint Statistical Meetings of the American Statistical Association, Survey Research Methods Section. Chicago, II.

Wolter, K.M. (1986). Some Coverage Error Models for Census Data. Journal of the American Statistical Association, 81, 336–346.

Zaslavsky, A., and Wolfgang, G. (1993). Triple System Modeling of Census, Post-Enumeration Survey, and Administrative-List Data. Journal of Business and Economic Statistics, 11, 279–288.

Received March 1999
Revised December 2000