Survey Methodology, December 2000
Vol. 26, No. 2, pp. 139-152
Statistics Canada, Catalogue No. 12-001

On the Validity of Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data

Paul P. Biemer and John M. Bushery [1]

Abstract

The primary goal of this research is to investigate the validity of Markov latent class analysis (MLCA) estimates of labor force classification error and to evaluate the efficacy of MLC analysis as an alternative to traditional methods for evaluating data quality. We analyze interview data from the Current Population Survey (CPS) for the first three months of each of three years – 1993, 1995, and 1996 – and conduct an additional analysis of the CPS unreconciled reinterview data for approximately the same time periods. The reinterview data provides another approach for estimating CPS classification error that, when compared with the MLC estimates, helps to address the validity of the MLCA approach. Five dimensions of MLCA validity are addressed: (a) model diagnostics, (b) model goodness of fit across three years of CPS, (c) agreement between the model and test-retest reinterview estimates of response probabilities, (d) agreement between the model and test-retest reinterview estimates of inconsistency, and (e) the plausibility of patterns of classification error. In addition, we consider the robustness of the MLCA estimates to violations in the Markov assumption. Our analyses provides no evidence to question the validity of the MLC approach. The method performed well in all five validity tests.

Key Words: Panel surveys; Nonsampling error; Unemployment; Data quality.

1. Introduction

The Current Population Survey (CPS) is a household sample survey conducted monthly by the U.S. Bureau of the Census to provide estimates of employment, unemployment, and other characteristics of the general U.S. labor force population. National estimates from the CPS of the size, composition, and changes in the composition of the labor force are published each month by the U.S. Bureau of Labor Statistics in Employment and Earnings. The CPS labor force estimates comprise one of the Nation's key economic indicators; since 1942, the Federal government has used the CPS data series to monitor month-to-month and year-to-year changes in labor force participation.

Given the importance of the CPS data series to public policy, there have been numerous evaluations of the accuracy of the data. For example, since the early 1950s, the Census Bureau has conducted the CPS Reinterview Program to evaluate the quality of the labor force data. The program involves drawing a small subsample (less than 5 percent) of the CPS respondents and re-asking some of the questions asked in the original interview – particularly the labor force questions. Until 1994, about one fourth of the sample received an unreconciled reinterview and three fourths received a reconciled reinterview. The reconciled reinterview component, which was used primarily for interview quality control purposes, was discontinued in 1994 due to concerns about the quality of the data. However, the unreconciled reinterview continues today and is used to estimate the test-retest reliability (or response consistency). Forsman and Schreiner (1991) provide a detailed description of the CPS Reinterview Program.

Several papers prepared by researchers outside the Census Bureau analyze the CPS Reinterview Program data to estimate the classification error in the CPS (*cf.* Sinclair and Gastwirth 1996, 1998; Biemer and Forsman 1992; Chua and Fuller 1987; Poterba and Summers 1986; Abowd and Zellner 1985). Recently, Poterba and Summers (1995) used data from the CPS Reinterview Program to estimate the CPS classification error rates and to evaluate the impact of classification error on labor market transition rates. As in the 1986 paper, their more recent analysis is based on the assumption that the CPS reinterview reconciliation process yields data which may be considered as the truth. Abowd and Zellner (1985) took similar approach.

Several authors (*viz.*, Sinclair and Gastwirth 1996, 1998; Biemer and Forsman 1992; Forsman and Schreiner 1991; Schreiner 1980) question the assumption that reconciled reinterview yields true values. They provide considerable evidence that the reinterview data are subject to substantial classification errors. In fact, this realization was responsible for the Census Bureau's decision to eliminate the reconciled reinterview portion of the CPS Reinterview Program in 1994.

As an alternative to the infallibility assumption, Chua and Fuller (1987) and Fuller and Chua (1985) apply a type of latent structure model to the CPS reconciled reinterview data to estimate the CPS response probabilities. For model identifiability, they impose tight restrictions on the response probabilities, forcing the bias due to classification error to be zero for both interview and reinterview. In addition, they assume independent classification errors for the interview and reinterview (referred to as the ICE assumption in the literature) and across the months in sample. The ICE

1. Paul P. Biemer, Research Triangle Institute, Research Triangle Park, NC 27709; John M. Bushery, Bureau of Transportation Statistics, Washington, DC 20590, U.S.A.

assumption is a limitation of their analysis because evidence in the literature suggests that the assumption may not hold for the CPS (see, for example, O'Muircheartaigh 1991, and Singh and Rao 1995). Consequently, response probabilities estimated using the Chua and Fuller approach may be biased.

Sinclair and Gastwirth (1996) and Sinclair and Gastwirth (1998) apply a latent class modeling approach to the CPS interview-reinterview data using model restrictions originally proposed by Hui and Walter (1980) for medical diagnostic testing. Using the interview-reinterview data cross-classified by sex, Sinclair and Gastwirth assume that classification error probabilities are equal for males and females while labor force participation rates differ for these groups. Since the model parameters consume all the available degrees of freedom for parameter estimation, no residual degrees of freedom are available to test model lack-of-fit. Consequently, their analysis does not directly address whether these model assumptions hold for the CPS data.

In an examination of the determinants of rotation group bias, Shockey (1988) also applies latent class analysis to the CPS. His analysis suggests that the rotation group bias problem first reported by Bailar (1975) may be caused by response error arising from the interview administration. Shockey did not use reinterview data but rather relied on confirmatory factor analytic methods to support his claims. The sizes of his error rates were much larger than those reported by other authors which may be an indication of model bias. Unfortunately, like Sinclair and Gastwirth, Shockey's data set is not adequate to test fully the assumptions of the model be used.

The method Markov latent class analysis, a promising approach for estimating the classification error in panel survey data, previously has not been applied to the CPS. This method takes advantage of the repeating nature of panel surveys to extract information on classification error directly from the interview data. The MLCA model is really a combination of two models: a latent Markov chain model representing the month to month transitions among the true labor force classifications and a classification error model representing the deviations from the true and observed labor classifications.

Because MLCA takes advantage of the repeating nature of panel surveys to extract information on classification error directly from the interview data, it does not require external, infallible measurements or remeasurements obtained by reinterview methods. In that regard, the method offers some advantages over both the Census Bureau's traditional methods and the methods of Chua and Fuller, Abowd and Zellner, Porterba and Summers, and Sinclair and Gastwirth for evaluating survey data quality in surveys. In many panel surveys, reinterviews are not feasible due to budget constraints, field work complexity, and respondent burden. MLCA may be the only way to assess the measurement error in these surveys. For panel surveys, such as the CPS, where reinterview data are available, the reinterview and MLCA methods offer alternative analytical approaches for evaluating classification error. For example, as in the present analysis, MLCA can be used to model and test the traditional reinterview analysis assumptions. Further, MLCA analysis provides a statistical framework for combining the panel data and reinterview data to obtain even more information about classification error (van de Pol and Langeheine 1997).

Another advantage of MLCA is the potential for incorporating the entire panel data set into the estimates of classification error rather than only the relatively small sample selected for reinterview. As a result, a number of data quality issues for panel surveys that previously could not be explored for lack of data may now be tractable.

This paper reports our findings regarding the utility of the MLCA modeling approach for evaluating labor force classification error in the CPS. Software for fitting a wide variety of MLCA and other latent class models is available from several sources. The software employed in our analysis is ℓEM (Vermunt 1997), which can fit a large class of log-linear models with or without latent variables. The flexibility and generality of this software allow the measurement error analyst to test a considerable range of classification error models and to explore hypotheses regarding the causes and correlates of classification error.

In the next section, we describe the MLCA model and estimation methodology and its theoretical underpinnings. In section 3, we develop the MLCA methodology for the CPS application, fit a series of models to the CPS, and examine the fit of these models. In this section, we also produce estimates of classification error based upon the best MLCA model. In section 4, we conduct a number of tests of the validity of the MLCA estimates including a comparison of the MLCA estimates with those from new interview-reinterview analysis. Finally, in section 5, we summarize our findings and make recommendations regarding the utility of the MLCA method for future evaluations of labor force classification error.

2. Markov Latent Class Analysis for Three Time Periods

Markov latent class models were first proposed by Wiggins (1973) and refined by Poulsen (1982). Van de Pol and de Leeuw (1986) established conditions under which the model is identifiable and gave other conditions of estimability of the model parameters. In this section, we develop the MLCA model in the context of the CPS and suggest other applications and its generalizations.

Let the CPS target population be divided into L groups (such as age, race, or sex groups) and let the variable G be the label for group membership. For example, $G_i = 1$ if the i^{th} population member is in group 1, $G_i = 2$ for group 2 and so on. Let X_{gi}, Y_{gi}, and Z_{gk} denote the true labor force classifications for the i^{th} person in group

$G = g$ (for $g = 1, ..., L$ and $i = 1, ..., n_g$) where X_{gi} is defined as

$$X_{gi} = \begin{cases} 1 & \text{if person } (g, i) \text{ is employed in time period 1} \\ 2 & \text{if person } (g, i) \text{ is unemployed in time period 1} \\ 3 & \text{if person } (g, i) \text{ is not in the labor force in time period 1} \end{cases}$$

with analogous definitions for Y_{gi} and Z_{gi} for periods 2 and 3 respectively. Let $\pi_{x, y, z|g}$ denote $\Pr(X = x, Y = y, Z = z | G = g)$, let $\pi_{y|g, x}$ denote $\Pr(Y = y | X = x, G = g)$ and let $\pi_{z|g, y, x}$ denote $\Pr(Z = z, Y = y, X = x | G = g)$. Then, the probability that an individual in group g has labor status x in period 1, y in period 2, and z in period 3 is

$$\pi_{x, y, z|g} = \pi_{x|g}\pi_{y|g, x}\pi_{z|g, x, y}. \quad (1)$$

Finally, under the first order Markov assumption, a necessary condition for model identifiability (see Van de Pol and de Leeuw 1986), we assume

$$\pi_{z|g, x, y} = \pi_{z|g, y} \quad (2)$$

i.e., at period 3, the true status of an individual does not depend on the period 1 status, once the period 2 status is known. An alternate interpretation is that the current status, given the prior period's status, does not depend upon the prior period's transition.

One can conceive of a number of scenarios where the Markov assumption may not hold for monthly labor force status. The assumption would be violated, for example, if individuals who are unemployed in period 2 are more likely to be unemployed in period 3, given they were also unemployed in period 1. The group of people unemployed in period 2 and period 1 probably includes a higher proportion of chronically unemployed people than the group unemployed in period 2, not period 1. That group (unemployed period 2 but not in period 1) likely contains a higher proportion of people temporarily out of work while changing jobs.

However, the validity of this assumption cannot be adequately explored using the observed data because the data are distorted to some unknown extent by the presence of classification errors. At least two methods for assessing the validity of the Markov assumption for panel data are available. Van de Pol and de Leeuw (1986) suggest a method based upon four waves of panel data that substitutes a second order Markov restriction for the first order restriction in (2). Another method, suggested by van de Pol and Langeheine (1997), uses a combination of labor force panel data and the reinterview data at each time period. Neither of these methods was employed in this paper to test the MLCA assumption directly. Instead, we assessed the overall validity of the MLCA estimates using the methods discussed in section 3.2 below. In section 3.6 we provide some results from a simulation study to illustrate the robustness of the MLCA estimates of classification error to violations of the Markov assumption.

Now, consider the observed labor force classifications from the CPS denoted by A_{gi}, B_{gi}, and C_{gi} for periods 1, 2, and 3, respectively, where

$$A_{gi} = \begin{cases} 1 & \text{if person } (g, i) \text{ is classified as employed in time period 1} \\ 2 & \text{if person } (g, i) \text{ is classified as unemployed in time period 1} \\ 3 & \text{if person } (g, i) \text{ is classified as NLF in time period 1} \end{cases}$$

with analogous definitions for the response indicators, B_{gi}, and C_{gi} for periods 2 and 3, respectively. Using an extension of the notation established above, we denote the response probabilities in each of these classifications as $\pi_{a|g, x} = \Pr(A = a | G = g, X = x)$, with analogous definitions for $\pi_{b|g, y}$ and $\pi_{c|g, z}$. Thus, $\pi_{a=1|g, x=2}$ is the probability that the CPS classifies a person in group g as employed $(A = 1)$ when the true status is unemployed $(X = 2)$. Likewise, $\pi_{a=2|g, x=2}$ is the probability that the CPS correctly classifies a person in group g as unemployed.

Finally, we assume

$$\pi_{a, b, c|g, x, y, z} = \pi_{a|gx}\pi_{b|gy}\pi_{c|gz} \quad (3)$$

or that classification error in the observed labor forces status is independent across the three months. This assumption, referred to as the local independence assumption, has been investigated for the CPS by Meyers (1988) in his review of the Abowd and Zellner (1985) estimation approach. Meyers concluded that the assumption "seems a reasonable approximation." Singh and Rao (1995), who studied the robustness of the assumption under a number of labor force population scenarios, reached a similar conclusion. Van de Pol and Langeheine (1997) modeled the joint distribution of panel data and reinterview data using latent class models to test for local independence for various types of labor force transitions. They found some evidence that people who change labor force status have lower reliability than those who do not, however the effect was quite small. Therefore, we shall also assume (3) without attempting any further investigation of its validity in this paper.

The CPS labor force classifications for each month of the first quarter of the year are the outcome variables in our analysis. Let A, B, and C denote the observed classifications and let X, Y, and Z denote the (unobserved) true classifications for January, February, and March, respectively. Let G denote some grouping (or stratification) variable to be defined later in the analysis. Under these assumptions, we can write the probability for classifying a CPS sample member in cell (g, a, b, c) of the *GABC* table as follows:

$$\pi_{g,a,b,c} = \sum_{x,y,z} \pi_g \pi_{x|g} \pi_{a|g,x} \pi_{y|x,g} \pi_{b|y,g} \pi_{z|g,y} \pi_{c|g,z}. \quad (4)$$

Extensions to more than one grouping variable are straightforward.

Under multinomial sampling, the likelihood function for the $GABC$ table is

$$\Pr(GABC) = k \prod_{g,a,b,c} \pi_{g,a,b,c}^{n_{gabc}} \quad (5)$$

where k is the multinomial constant and $\prod$ denotes the product of the terms over the subscripts g, a, b, and c. Under the assumptions made previously, the model parameters are estimable using maximum likelihood estimation methods. Van de Pol and de Leeuw (1986) provides the formula for applying the E-M algorithm to estimate the parameters of this model and describes the conditions for their estimability. The ℓEM software, applied to the CPS data sets in the next section, implements these methods.

3. Application to the CPS

3.1 Notation

Part of our evaluation of the MLCA approach will compare the MLCA estimates of classification error with estimates derived from the analysis of interview-reinterview data. Using the notation in the previous section, let A and A' denote the labor force classification for the original and reinterview, respectively, and define $\pi_a = \Pr(A = a)$ and $\pi_{a'} = \Pr(A' = a')$. Let AA' denote the observed interview-reinterview $K \times K$ cross-classification table and let $\boldsymbol{\pi}^{AA'|X}$ denote the $K \times K$ matrix of cell probabilities, $\Pr(A = a, A' = a' | X = x)$. If we assume that $\pi'_{aa} = \Pr(A = a, A' = a' | X = x) = \pi^2_{a|x}$, referred to in the literature as the assumption of parallel measures (Bohrnstedt 1983), then

$$\boldsymbol{\pi}^{AA'|X} = \boldsymbol{\pi}^{A|X} (\boldsymbol{\pi}^{A|X})^T \quad (6)$$

where $(\boldsymbol{\pi}^{A|X})^T$ denotes the transpose of vector of conditional probabilities, $\boldsymbol{\pi}^{A|X}$.

Let $\boldsymbol{\pi}^X$ denote the K-vector of true classification probabilities. Then

$$\boldsymbol{\pi}^{AA'} = \boldsymbol{\pi}^{AA'|X} \boldsymbol{\pi}^X \quad (7)$$

i.e., the probability of the observed interview-reinterview classification table, $\boldsymbol{\pi}^{AA'}$, is equal to the product of the matrix of conditional response probabilities, $\boldsymbol{\pi}^{AA'|X}$, and the vector of true classification probabilities, $\boldsymbol{\pi}^X$.

As described in the previous section, the MLCA of the CPS longitudinal data will provide maximum likelihood estimates of $\boldsymbol{\pi}^{A|X}$ and $\boldsymbol{\pi}^X$, allowing the estimation of $\boldsymbol{\pi}^{AA'}$ via (6) and (7). We can estimate the test-retest reliability, R, for any labor force category by applying the usual estimation methods (see, for example, Bohrnstedt 1983) to this estimate of $\boldsymbol{\pi}^{AA'}$. For our analysis, we compute the index of inconsistency, $I = 1 - R$, which is the traditional reliability measure for CPS labor force data (see U.S. Bureau of the Census 1985). Let I_a, denote the index of inconsistency for category $A = a$. Then an estimator of I_a is

$$\frac{gdr}{2\hat{\pi}_a(1 - \hat{\pi}_a)} \quad (8)$$

where gdr is the gross difference rate defined by

$$gdr_a = 2 \sum_{a \neq a'} \hat{\pi}_{a,a'} \quad (9)$$

and where $\hat{\pi}_a$ and $\hat{\pi}_{a,a'}$ denote latent class estimates of π_a and $\pi_{a,a'}$, respectively.

U.S. Bureau of the census (1985, 88-91) provides the formulas for standard errors as well as an aggregate measure of inconsistency for all K categories combined, referred to as the aggregate index of inconsistency, I_{AG}. The aggregate index is a question-level measure of unreliability equal to $1 - \kappa$ (Hess, Singer and Bushery 2000) where κ is Cohen's kappa reliability measure (Cohen 1960) and is a weighted average of the category-level indexes.

Finally, given an estimate of $\boldsymbol{\pi}^X$ we can estimate the K-vector of measurement biases, denoted by $\boldsymbol{\beta}_A$, associated with the K categories of A using the identity

$$\boldsymbol{\beta}_A = \boldsymbol{\pi}^A - \boldsymbol{\pi}^X. \quad (10)$$

3.2 Assessing the Validity of the MLCA Methodology

The primary objective of this paper is to assess the validity of the MLCA approach. Previous research in the measurement of CPS classification error has not fully addressed the validity of the estimation approaches used (Meyers 1988). We hope to determine whether the MLCA approach is informative and useful for studying classification error in the CPS. In particular, we aim to determine whether the model estimates of error probabilities, $\boldsymbol{\pi}^{A|X}$, reflect the actual levels of error in the CPS labor force classifications. Unfortunately, for the reasons mentioned previously, no generally accepted gold standard exists for assessing the accuracy of the CPS (see, for example, Sinclair and Gastwirth 1996, 1998, Biemer and Forsman 1992, and Schreiner 1980). Consequently, estimating the bias of MLCA estimates is not possible.

In what follows, we will investigate the validity of the MLCA estimates of CPS classification error using five criteria:

1. **Model diagnostics.** A necessary condition for model validity is that the model is plausible (*i.e.*, the assumptions are reasonable and are consistent with reality) and fits the data adequately. We use the traditional chi-square goodness of fit criteria and other diagnostic measures of model fit to assess the adequacy of the model specification and the degree to which the data are consistent with the model.

2. **Model Goodness of Fit Across Years of CPS.** An often-used technique for model validation is to assess the fit of the model for data that are independent of the data used for model building (see, for example, Kleinbaum, Kupper and Muller 1988, 330). This method is useful for avoiding model over-parameterization and data-driven (rather than theory-driven) model selection. In the present study, fitting the same model to data for each year separately is a form of this independent model verification technique. Model agreement across the years would tend to support the validity of the model structure. This method has a difficulty in the present application. After 1993, the CPS paper and pencil questionnaire was redesigned for Computer Assisted Personal Interview (CAPI) administration, so the magnitudes of the response errors may have changed after 1993. However, if the primary sources of response error in the CPS have not changed with the redesign, a model structure that adequately describes the error for 1993 should also describe the error for 1995 and 1996.
3. **Agreement of the MLCA Estimates and the Hui-Water Test-Retest Estimates of Response Probabilities.** The Hui-Walter (H-W) method (Hui and Walter 1980) for estimating CPS response probabilities uses unreconciled reinterview data (Sinclair and Gastwirth 1996; 1998). Although the MLCA and H-W methods both use latent class models, the model assumptions are very different. For example, the H-W method does not require the Markov assumption for model identifiability. Further, in this research, the data inputs to the H-W method are independent of those used for the MLCA method. Close agreement between the two sets of estimates supports the validity of both methods, while poor agreement suggests that at least one of the approaches is not valid. Strong agreement between the MLCA and H-W estimates also lend some assurance that the MLCA estimates of response probabilities are relatively robust to possible violations of the Markov assumption.
4. **Agreement of Model and Test-Retest Estimates of the Index of Inconsistency.** This criterion is similar to Criterion 3 because it compares estimates derived from MLCA with estimates based upon unreconciled reinterview data. However, this analysis does not rely on the validity of the Hui-Walter estimation methodology to assess MLCA estimation validity. Instead we use the MLCA estimates of classification error to compute estimates of the index of inconsistency using (7) to (9). We compare these estimates of reliability directly to the estimates of reliability from the CPS Reinterview Program, obtained from unreconciled reinterview data. Good agreement between the Reinterview and MLCA estimates supports the validity of both methods, while poor agreement suggests that at least one of the approaches is not valid.
5. **Plausibility of Patterns of Classification Error**. Finally, the plausibility (or face validity) of the response probability estimates can also provide a test of validity. For example, it seems implausible that proxy responses to labor force questions should be more accurate than self-responses. Other patterns of classification error can also be reviewed and evaluated for plausibility. To the extent that the model estimates seem plausible, the face validity of the estimates is supported.

In the next section, we discuss our MLCA modeling results in the context of these criteria for validity. We begin with a description of the CPS data sets and the results of the model selection process.

3.3 The CPS Data Sets

In 1994, in conjunction with the implementation of computer assisted personal interviewing (CAPI), the CPS underwent a major redesign and a restructuring of the questions used to determine labor force status. Rothgeb (1994) provides a description of the CPS redesign. As a result of these improvements, we expect to see a difference (specifically a reduction) in classification error for the post-1994 CPS relative to 1993. Although not a primary objective of this research, we compared the error in the CPS before and after the redesign. We tested the MLCA approach for three years of the CPS – 1993, 1995, and 1996 – because the CPS unreconciled reinterview data were readily available for these time periods.

The CPS households are interviewed for four consecutive months, drop out of the survey for eight months, and then re-enter to be interviewed for a second series of four consecutive months. MLCA requires at least three consecutive interviews for identifiability of the model parameters. We had a choice of data sets which included all persons interviewed in three or four consecutive months of the CPS. Since using four months of data would reduce the sample size for the analysis by half, we chose to focus the analysis on three consecutive months – January, February, and March – for all three years of data. Nonresponse cases and cases where the whole household changed in one or more of the three months were excluded from the analysis.

The simplest MLCA model specifies that the response probabilities, $\pi_{a|x}, \pi_{b|y}$, and $\pi_{c|z}$, and the transition probabilities, $\pi_{y|x}, \pi_{z|y}$ are the same for all persons in the target population (referred to as homogeneity). However, our preliminary analysis (Biemer, Bushery and Flanagan 1997) indicated that response and transition probabilities where not homogeneous. To account for this heterogeneity, we explored a number of covariates and stratification variables for inclusion in the models, including: gender, education, mode of interview, proxy/self-response, and race. Of the

those considered, a variable derived from the CPS proxy/self response indicator best accounted for population heterogeneity. This variable, denoted by P, is defined as follows:

$$P = \begin{cases} 1 & \text{if all three interviews are conducted by self-response (SELF)} \\ 2 & \text{if two of the three interviews are conducted by self-response (MOSTLY SELF)} \\ 3 & \text{if two of the three interviews are conducted by proxy response (MOSTLY PROXY)} \\ 4 & \text{if all three interviews are conducted by proxy response (PROXY)} \end{cases}$$

Note, we now use P to represent the grouping variable, in place of G, which we used in section 2. Based upon previous research (for example, O'Muircheartaigh 1991), we expect that the Self group ($P = 1$) to have less classification error than the Proxy group ($P = 4$). We test this hypothesis as part of the estimate plausibility criterion (criterion 4 above).

The sample sizes for the three data sets used in our analysis are

1993: 45,291 persons
1995: 49,347 persons
1996: 41,751 persons

For 1993, approximately one-third of the sample is in the Self group, approximately one-fourth in the Proxy group, and the remaining sample members are distributed approximately equally between the Mostly Self and Mostly Proxy groups. For 1995 and 1996, slightly more sample members (one-third rather than one-fourth) are in the Proxy group.

3.4 Fitting the MLCA Models

To fit an MLCA model with a single grouping variable, P, the input data set was a 4 3 3 3 table of cell counts defined by the cross-classification of P A B C, where A, B, and C are the labor force classifications for January, February, and March, respectively.

The ℓEM software and other software packages for fitting MLCA models assume simple random sampling, so the complex survey design of the CPS cannot be modeled exactly. It is possible to account for the unequal probability sampling structure of the CPS through the use of weighted and rescaled cell counts rather than the raw cell totals (Clogg and Eliason 1985). However, using unweighted data for the MLCA analysis affords two important advantages. First, we can compare the MLCA estimates with estimates from the previously cited studies on CPS classification error, all of which used unweighted data. Second, the CPS reinterview data used to assess Criteria 3 and 4 are unweighted and weights are not available. Consequently, at least part of the analysis requires unweighted data; using weighted data for the other criteria could produce spurious inconsistencies in the results.

To investigate the validity of inferences to the total population using unweighted analysis, we estimated classification errors from both weighted and unweigted data and observed that the classification error estimates expressed as proportions were virtually identical, differing only at the third decimal place. Thus, the results we report below using unweighted cell counts are appropriate for inference beyond the CPS sample to the total population.

Another consideration in using unweighted analysis is the estimation of standard errors. Since they are computed using simple random sampling assumptions, the ℓEM standard error estimates may be understated as a result of ignoring the clustering effects in the CPS sample. To approximately account for this, we can multiply the ℓEM variances by a design effect computed from the CPS labor force estimates. U.S. Bureau of the Census (2000, 14-9) indicates that the design effects for the CPS labor force estimates do not exceed 1.3 and thus multiplying the ℓEM standard errors by (1.3)½ should inflate the standard errors sufficiently to account for clustering. An equivalent approach is to use a 3 percent rather than a 5 percent level of significance in declaring the difference between two estimates to be statistically significant. This latter strategy will be employed in the forthcoming analysis as appropriate. We believe this produces a conservative test since the CPS design effect reflects the increase in variance due to both sample clustering and unequal weighting, while only clustering effects are present our unweighted estimates.

Table 1 shows the results of fitting a sequence of increasingly complex MLCA models for each of the three data sets. The Base Model is the simplest MLCA model and specifies that transition probabilities and response probabilities are homogeneous (*i.e.*, do not differ by group, P) and stationary (*i.e.*, are the same for all three months). This model may be written as

$$\pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \pi_{x|g} \pi^3_{a|x} \pi^2_{y|x,} \qquad (11)$$

which is obtained from (4) by imposing the constraints

$$\pi_{z|yp} = \pi_{y|xp} = \pi_{y|x} \qquad (12)$$

and

$$\pi_{a|xp} = \pi_{b|yp} = \pi_{c|zp} = \pi_{a|x} \qquad (13)$$

for all p.

For Model 1 we relax constraint (12) to

$$\pi_{z|yp} = \pi_{y|xp} \text{ for } p = 1, ..., 4 \qquad (14)$$

and thus allow transitions from January to February and February to March to vary by Self/Proxy Group, P. For Model 2, we further relax constraint (12) to

$$\pi_{z|xp} = \pi_{y|x} \text{ and } \pi_{z|yp} = \pi_{z|y} \qquad (15)$$

Table 1
Model Diagnostics for Alternative MLCA Models by Year

1993 Data	*df*	npar[1]	L^2	*p*-value	BIC	*d*
Base Model: Homogeneous and stationary transitions and response probabilities	90	17	645	0	-320	0.048
Model 1: Nonhomogeneous transitions	84	23	632	0	-269	0.047
Model 2: Non-stationary transitions	66	41	99	0.006	-609	0.01
Model 3: Nonhomogeneous and non-stationary transitions	42	65	64	0.016	-386	0.01
Model 4: Nonhomogeneous and non-stationary transitions and nonhomogeneous response probabilities	24	83	23	0.501	-234	0
1995 Data	*df*	npar[1]	L^2	*p*-value	BIC	*d*
Base Model: Homogeneous and stationary transitions and response probabilities	90	17	697	0	-275	0.044
Model 1: Nonhomogeneous transitions	84	23	668	0	-240	0.043
Model 2: Non-stationary transitions	66	41	146	0	-567	0.01
Model 3: Nonhomogeneous and non-stationary transitions	42	65	82	0	-372	0.01
Model 4: Nonhomogeneous and non-stationary transitions and nonhomogeneous response probabilities	24	83	25	0.41	-234	0
1996 Data	*df*	npar[1]	L^2	*p*-value	BIC	*d*
Base Model: Homogeneous and stationary transitions and response probabilities	90	17	632	0	-325	0.045
Model 1: Nonhomogeneous transitions	84	23	585	0	-308	0.044
Model 2: Non-stationary transitions	66	41	159	0	-543	0.01
Model 3: Nonhomogeneous and non-stationary transitions	42	65	82.6	0	-364	0.01
Model 4: Nonhomogeneous and non-stationary transitions and nonhomogeneous response probabilities	24	83	39.3	0.026	-216	0

[1] Note that "npar" refers to the number of parameters in the model.

for all p. Model 3 relaxes both the homogeneity and stationarity constraints for transition probabilities so that $\pi_{y|xp} \quad \pi_{z|yp}$. Thus this model allows transition probabilities to vary by group and by month. However, response probabilities are still constrained to be equal across groups and months.

Model 4 is the most general, identifiable model we considered. Model 4 allows the January-February and February-March transition probabilities to vary independently across the four proxy/self groups. This model further species that the response probabilities are the same for January, February, and March, but may vary across the four proxy/self groups. We obtained this model from Model 3 by relaxing the constraints specifying homogeneous response probabilities; *i.e.*, by relaxing constraint (13) to

$$\pi_{a|xp} = \pi_{b|yp} = \pi_{c|zp} \tag{16}$$

for all p. Under these constraints, (4) can be written as

$$\pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \pi_{x|p} \pi_{y|xp} \pi_{z|py} (\pi_{a|p,x})^3.$$

In Table 1, we show the basic fit statistics for all five models for all three years. Column 4 of the table provides L^2, the usual likelihood ratio chi-square statistics (see Agresti 1990, 48), and column 5 the corresponding p-value. A p-value of 0.05 or greater is the usual criterion for adequate model fit. However, due to the large sample sizes in our analysis, requiring a p-value this large could result in model over fitting. We consider a p-value as small as 0.01 to be acceptable. The BIC measure in the table is defined as

$$\text{BIC} = L^2 - (\log N)\, df$$

where N is the total sample size and df is the degrees of freedom for the model. The BIC essentially summarizes the tradeoff between model fit (L2) and model parsimony (df). Since small values of the BIC are favorable, we will regard the model with the smallest BIC as best with respect to goodness of fit and parsimony. Liu and Dayton (1997) discuss this approach for latent class models.

Finally, the dissimilarity index (d) is the proportion of observations that would have to change cells for the model to fit perfectly. As rule of thumb, models having $d \le 0.05$ (*i.e.*, 5 percent model error) are considered to fit the data well (Vermunt 1997).

For each year of data, Model 4 is the only model to provide an acceptable fit when the p-value criterion is

considered. Model 4 is also plausible from a response theory perspective since it postulates that classification error varies by self/proxy group. This, as we have said, is consistent with the survey methods literature (see for example, O'Muircheartaigh 1991 and Moore 1988). The dissimilarity index, d, for the model is 0.3 percent, which indicates a very good model fit. Thus, we use Model 4 to generate the estimates of labor force classification error.

3.5 Estimation of Classification Error

Table 2 shows estimates of the response probabilities from Model 4 for Employed, Unemployed, and Not in the Labor Force. For the true Employed and those truly Not in the Labor Force, the probability of a correct response is quite high: at least 98 percent for Employed and 97 percent for Not in the Labor Force. However, for the true Unemployed, the probability of a correct response varies across years and groups from approximately 68 percent to approximately 86 percent. As expected, the Self group has the highest probability of a correct response (statistically significant for 1993 and 1996). The Mostly Self group also displays a tendency for a higher probability of a correct response than the Mostly Proxy group, but the difference is not statistically significant.

A surprising result from Table 2 is the direction of the difference in reporting accuracy for the true unemployed across the three years. Recall that the CPS interview questionnaire was redesigned in 1994 to increase reporting accuracy. However, these results suggest that reporting accuracy is higher for 1993 (the year prior to the major redesign) than for 1995 and 1996 (the years following the redesign). In 1993, the probability of a correct response was 81.8 percent, compared with 76.1 percent and 74.4 percent for 1995 and 1996, respectively (p - value < 0.001). This effect may be a consequence the redesign or may reflect actual changes in the population, or both. Investigations are currently underway at the Census Bureau to understand the causes underlying these results.

Table 2
Estimated Labor Force Classification Probabilities by Group and Year (Standard Errors are in parentheses)

Observed	Group	True Employed			True Unemployed			True NLF		
		1993	1995	1996	1993	1995	1996	1993	1995	1996
Emp	Self	98.90	98.98	99.03	4.24	5.66	5.78	1.05	0.67	0.92
		(0.12)	(0.12)	(0.13)	(0.96)	(1.28)	(1.52)	(0.13)	(0.10)	(0.13)
	Mostly Self	98.91	98.61	98.86	8.11	7.66	12.20	1.73	1.59	0.93
		(0.18)	(0.21)	(0.22)	(1.89)	(2.50)	(2.96)	(0.27)	(0.30)	(0.24)
	Mostly Proxy	98.62	98.56	98.76	11.19	12.86	10.18	1.68	0.95	1.50
		(0.18)	(0.22)	(0.20)	(1.96)	(2.97)	(2.73)	(0.27)	(0.25)	(0.30)
	Proxy	98.66	98.63	98.66	6.60	7.77	8.30	1.56	1.50	1.30
		(0.16)	(0.13)	(0.15)	(1.43)	(1.44)	(1.62)	(0.20)	(0.17)	(0.18)
	Total	98.77	98.73	98.83	7.06	7.86	8.57	1.41	1.11	1.13
		(0.11)	(0.11)	(0.11)	(0.70)	(0.90)	(1.00)	(0.11)	(0.11)	(0.11)
Unemp	Self	0.33	0.39	0.41	85.92	80.52	79.66	0.48	0.35	0.40
		(0.08)	(0.08)	(0.09)	(1.62)	(2.04)	(2.44)	(0.11)	(0.10)	(0.12)
	Mostly Self	0.28	0.50	0.28	82.68	71.71	73.60	1.03	0.91	0.79
		(0.12)	(0.13)	(0.15)	(2.48)	(3.40)	(3.71)	(0.23)	(0.21)	(0.28)
	Mostly Proxy	0.37	0.54	0.37	76.54	68.09	72.74	0.81	0.90	1.53
		(0.11)	(0.15)	(0.13)	(2.55)	(3.37)	(3.70)	(0.22)	(0.28)	(0.26)
	Proxy	0.34	0.55	0.39	80.09	77.12	71.63	0.95	0.94	1.18
		(0.10)	(0.09)	(0.09)	(2.19)	(2.18)	(2.49)	(0.19)	(0.17)	(0.20)
	Total	0.34	0.49	0.37	81.81	76.09	74.42	0.75	0.69	0.87
		(0.11)	(0.11)	(0.11)	(0.90)	(1.21)	(1.21)	(0.11)	(0.11)	(0.11)
NLF	Self	0.77	0.63	0.55	9.84	13.82	14.56	98.47	98.98	98.68
		(0.10)	(0.08)	(0.09)	(1.39)	(1.76)	(2.11)	(0.17)	(0.14)	(0.18)
	Mostly Self	0.81	0.89	0.86	9.21	20.63	14.20	97.24	97.50	98.28
		(0.13)	(0.15)	(0.16)	(1.83)	(2.94)	(2.94)	(0.35)	(0.36)	(0.37)
	Mostly Proxy	1.01	0.90	0.87	12.27	19.05	17.08	97.52	98.15	96.96
		(0.14)	(0.15)	(0.15)	(1.97)	(2.58)	(3.16)	(0.35)	(0.38)	(0.40)
	Proxy	0.10	0.82	0.95	13.31	15.11	20.07	97.49	97.56	97.52
		(0.13)	(0.10)	(0.12)	(1.83)	(1.86)	(2.22)	(0.27)	(0.24)	(0.27)
	Total	0.89	0.78	0.79	11.13	16.04	17.00	97.84	98.20	98.00
		(0.11)	(0.11)	(0.11)	(0.90)	(1.21)	(1.21)	(0.11)	(0.11)	(0.11)

Table 3
Comparison of MLCA Estimates with Prior Published Estimates

Classification		MLCA	Chua & Fuller (1982 data)	Poterba & Summers (1981 data)	CPS Reconciled Reinterview (1977-1982)
True	Observed				
Employed	Emp	98.77 (1993)	98.66 (month 1)	97.74	98.78
		98.73 (1995)	98.65 (month 2)		
		98.73 (1996)			
	Unemp	0.34 (1993)	0.32 (month 1)	0.54	0.19
		0.49 (1995)	0.34 (month 2)		
		0.37 (1996)			
	NLF	0.89 (1993)	1.02 (month 1)	1.72	1.03
		0.78 (1995)	1.01 (month 2)		
		0.79 (1996)			
Unemp	Emp	7.06 (1993)	3.52 (month 1)	3.78	1.91
		7.86 (1995)	3.51 (month 2)		
		8.57 (1996)			
	Unemp	81.81 (1993)	88.27 (month 1)	84.76	88.57
		76.09 (1995)	88.23 (month 2)		
		74.42 (1996)			
	NLF	11.13 (1993)	8.21 (month 1)	11.46	9.53
		16.04 (1995)	8.16 (month 2)		
		17.00 (1996)			
NLF	Emp	1.41 (1993)	1.60 (month 1)	1.16	0.5
		1.11 (1995)	1.61 (month 2)		
		1.13 (1996)			
	Unemp	0.75 (1993)	1.19 (month 1)	0.64	0.29
		0.69 (1995)	1.24 (month 2)		
		0.87 (1996)			
	NLF	97.84 (1993)	97.21 (month 1)	98.2	99.21
		98.20 (1995)	97.15 (month 2)		
		98.00 (1996)			

The table indicates that misclassification of the unemployed as NLF is a bigger problem than misclassification as Employed. Averaging over all three years, approximately two thirds of the error in classifying the unemployed is misclassification as NLF. But the rates of both types of error are high.

Next, we compare our estimates of the CPS classification probabilities with similar estimates from the literature. In Table 3, the MLCA estimates for each of the three years are compared with estimates from Chua and Fuller (1987), Poterba and Summers (1995), and the CPS reconciled reinterview program. Again, the latter three sets of estimates rely on reinterview data while the MLCA estimates are produced directly from the CPS interview data. In general, the relative magnitude of the MLCA estimates across the labor force categories agrees with the previous estimates. The greatest differences occur for the true unemployed population. For this group, the estimates of response accuracy from the literature are three to seven percentage points higher than corresponding MLCA estimates for 1993, which is the time period that most closely corresponds to the comparison estimates.

One explanation for this difference is that the comparison estimates are biased upward as a result of correlations between the errors in interview and reinterview. Another explanation is that the MLCA estimates are biased downward as a result of the failure of the Markov assumption to hold. We suspect that both explanations may be true to some extent. However, the next section provides some evidence that failure of the Markov assumption likely has a small effect on estimates of classification error.

3.6 Robustness of MLCA to Non-Markov Labor Force Transitions

A number of authors have investigated the effects of current and previous employment status on future employment status (see, for example, Akerlof and Main 1980; Heckman and Borjas 1980; Lynch 1989, and Corak 1993). Heckman and Borjas show that examination of this issue is quite difficult due to selection biases, response error, and unobserved heterogeneity. These confounding influences may account for the inconsistent findings in the literature. For example, using data from the CPS, Akerlof and Main (1980) provide evidence that the probability of

future unemployment depends upon the number of previous unemployment spells experienced as well as the duration of those spells. However, in a study of male high school graduates, Heckman and Borjas (1980) found "no evidence that previous occurrences of unemployment or their duration affect future labor market behavior once we control for sample selection bias and heterogeneity bias." The results from the literature are also inconsistent and ambiguous regarding the extent to which the Markov assumption expressed in (2) may be violated for the CPS and other labor market surveys. Nevertheless, in this section, we attempt to provide at least a partial answer to question of how non-Markov labor force transitions affect MLCA estimates of classification error.

To investigate the effect of violations of the Markov assumption in (2) for the present application, we conducted a limited simulation study. To focus the investigation while simplifying the simulation framework, we considered latent structures involving only two classes or states at each time point: unemployed, denoted by X, Y, or $Z = 1$, and other (*i.e.*, employed or not in the labor force), denoted by X, Y, or $Z = 2$ with analogous definitions for the observed states A, B, and C. To create a population for the simulation, the latent probabilities $\pi_x, \pi_{y|x}$, and $\pi_{z|xy}$ and the response probabilities $\pi_{a|x} = \pi_{b|y} = \pi_{c|z}$ were specified to be consistent with the combined 1993, 1995, and 1996 data sets.

We then defined two parameters, λ_1 and λ_2 to be varied in the simulation, where

$$\lambda_1 = \frac{\pi_{z=1|x=2,y=1}}{\pi_{z=1|x=1,y=1}} \tag{17}$$

and

$$\lambda_2 = \frac{\pi_{z=1|x=2,y=2}}{\pi_{z=1|x=1,y=2}}. \tag{18}$$

Thus, λ_1 is the probability of being "unemployed" in March, given "unemployed" in February and "other" in January over the probability of being "unemployed" in March given "unemployed" in the two previous months. Consistent with the findings of Akerlof and Main (1980) who showed that the likelihood of remaining unemployed increases as the number of unemployment spells increases, we assume that $0 \le \lambda_1 \le 1$. Similarly, λ_2 is the probability of being "unemployed" in March, given "other" in the two previous months, over the probability of being "unemployed", given "other" in February and "unemployed" in January. Again, by Akerlof and Main, we assume $0 \le \lambda_2 \le 1$. Note that when $\lambda_1 = \lambda_2 = 1$, unemployment transitions from February to March are Markov.

The simulated data were generated to be completely consistent with a MLCA model having non-stationary transition probabilities when $\lambda_1 = \lambda_2 = 1$. We simulated failure of the Markov assumption by varying λ_1 and λ_2 between 0 and 1. To be consistent with the 1993-1996 data, we fixed the probability of a correct "unemployed" response $\pi_{a=1|x=1}$, at 0.80 and the probability of a correct "other" response $\pi_{a=2|x=2}$, at 0.99 in all simulations. In addition, the denominators of λ_1 and λ_2 were fixed to their values as determined from the combined 1993-1996 data while the numerators were computed from (17) and (18) using the values of λ_1 and λ_2 specified in each simulation run.

Table 4 summarizes the results of the simulation for $\lambda_1 = \lambda_2 = \lambda$ where λ is varied from 0.2 to 1.0 to steps of 0.2. Note that for $\lambda_1 = \lambda_2 = 1.0$, which corresponds to a Markov model, the estimated probabilities of correct response are exactly as specified. For smaller values of λ_1 and λ_2, the estimates become negatively biased and are most biased for the lowest value considered, 0.2. Nevertheless, the absolute biases due to non-Markov transitions probabilities are never more than 3 percentage points. The results in Table 4 are consistent with Bushery and Kindelberger (1999), who used a somewhat different approach to illustrate the same robustness property of the MLCA models for CPS data. Both studies suggest that failure of the Markov assumption to hold does not appear to be an important source of bias in estimating CPS classification error probabilities.

Table 4
Estimates of Correct Classification Under
Non-Markov Transitions (Cell entries are percentages)

	$\lambda_1 = \lambda_2 = \lambda$				
Pr (Correct)	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 1.0$ (Markov)
Pr ("unemp"| true "unemp") = $\pi_{a=1\|x=1}$	77.6	78.1	78.7	79.3	80
Pr ("other"| true "other") = $\pi_{a=2\|x=2}$	98.6	98.7	98.8	98.9	99

4. Comparing the MLCA and Unreconciled Reinterview Estimates

4.1 Hui-Walter Estimation

An alternative set of response probability estimates can be obtained from the CPS reinterview data using a type of latent class model first proposed by Hui and Walter (1980). Using the notation introduced above, let X denote the true labor force classification for some time point and let A and A' denote the interview and reinterview classifications, respectively. Let G denote a grouping variable defined as in (4). Consider the likelihood of the group interview reinterview table denoted by GAA'. Denote by $\pi_{gaa'}$ the probability of classifying an individual belonging to group g into cell (a, a') of the table. The model for $\pi_{gaa'}$ proposed by Hui and Walter is

$$\pi_{gaa'} = \sum_x \pi_g \pi_{x|g} \pi_{a|x} \pi_{a'|x}. \tag{19}$$

In this model, the parallel measures assumption for the interview and reinterview responses is relaxed and response probabilities for the two measures, *viz.* $\pi_{a|x}$ and $\pi_{a'|x}$, are estimated separately. The ICE assumption is made as a condition of identifiability. It is further assumed that $\pi_{a|x}$ and $\pi_{a'|x}$ do not depend upon the group variable, *G*, while the prevalence of employed, unemployed, and NLF, *i.e.* $\pi_{x|g}$, still depends upon *G*.

Sinclair and Gastwirth's (1996) analysis of CPS labor force classification error used Sex as the grouping variable and our analysis uses this grouping variable as well. Sinclair and Gastwirth confined their analysis to white males and females and two labor force categories: NLF and In the Labor Force. The latter category is the sum of our Employed and Unemployed categories. In our analysis, we consider sample members of all races and analyze the three category labor force classification used in the MLCA. Thus, the H-W analysis estimates 16 parameters for each year, which equals the number of degrees of freedom available from the $G \quad A \quad A'$ table, leaving no degrees of freedom to test model fit.

The ℓEM software was used to fit the H-W model to the interview and unreconciled reinterview data from three time periods that coincide with the three in our MLCA: pre-1994, 1995, and 1996. We attempted to restrict the analysis to only the first quarter of these time periods. Unfortunately due the small sample sizes, the estimates were quite unstable. Thus, it was necessary to use the reinterview data from all four quarters of these time periods. The pre-1994 data were collected from 1985 through 1988 using the unreconciled reinterview sample.

The results of this comparison of MLCA and H-W estimates are summarized in Table 5. The MLCA estimates are the same as those in the rows of Table 2 labeled "Total." The H-W estimates are the classification probabilities associated with the original interview, *i.e.*, measure A in (19). The table shows the comparison for all three years. Since the largest error rate in the MLCA occurred for the Unemployed, this category is of particular interest in the MLCA/H-W comparison.

Overall, the two sets of estimates show fairly good agreement. The years 1995 and 1996 exhibit no statistically significant differences (at the 5 percent level) between the MLCA and H-W estimates for the unemployed population. The pre-1994 estimates display significant differences; however, they may be explained by the fact that the pre-1994 reinterview data were from 1985 through 1988, rather than 1993. These differences will be explored further in the next section.

4.2 Comparison of Indexes of Inconsistency

As described in section 3.1, we compute estimates of the index of inconsistency for each time period using the MLCA model-based estimates of the response probabilities. Essentially, we estimate the expected interview-reinterview cross-classification table from the MLCA response probability estimates and then apply the formula for the index to this table as though the table were observed. A second expected interview-reinterview classification table can be estimated using the H-W response probability estimates. We then compared these two sets of estimates to the estimate of the index computed directly from the CPS reinterview data using traditional methods (U.S. Bureau of the Census 1985). Agreement of all the three estimates agree supports the validity of the three methods.

Table 5
Comparison of MLCA and H-W Model Estimates of CPS Response Probabilities by Year (Standards Errors are in Parentheses)

Classification		1993		1995		1996	
True	Observed	H-W	MLCA	H-W	MLCA	H-W	MLCA
Emp	Emp	99.3	98.8	99.5	98.7	99.6	98.8
		(0.3)	(0.1)	(0.7)	(0.1)	(0.1)	(0.1)
	Unemp	0.0	0.3	0.0	0.5	0.4	0.4
		(0.0)	(0.1)	(n/a)	(0.1)	(0.1)	(0.1)
	NLF	0.7	0.9	0.5	0.8	0.0	0.8
		(0.3)	(0.1)	(0.7)	(0.1)	(n/a)	(0.1)
Unemp	Emp	11.1	7.1	11.5	7.9	4.6	8.6
		(1.0)	(0.7)	(2.3)	(0.9)	(15.2)	(1.0)
	Unemp	74.3	81.8	67.9	76.1	67.6	74.4
		(2.7)	(1.1)	(6.1)	(1.3)	(11.1)	(1.4)
	NLF	14.7	11.1	20.6	16.0	27.9	17.0
		(2.9)	(0.9)	(6.5)	(1.2)	(5.3)	(1.2)
NLF	Emp	2.0	1.4	2.5	1.1	2.6	1.1
		(0.5)	(0.1)	(1.5)	(0.1)	(1.5)	(0.1)
	Unemp	1.2	0.8	0.5	0.7	0.0	0.9
		(0.3)	(0.1)	(0.6)	(0.1)	(n/a)	(0.1)
	NLF	96.8	97.8	97.0	98.2	97.4	98.0
		(0.6)	(0.1)	(1.6)	(0.1)	(1.1)	(0.1)

Table 6 shows the three methods estimates the index of inconsistency for all three time periods. As before, the Unemployed category is of particular interest because of its large error rate. Standard errors are not available for the MLCA or the H-W estimates of the index so formal tests of hypothesis are not possible. However, standard errors for the traditional estimates are provided which can be used as rough approximations of the standard errors for the H-W estimates.

Overall, both the general patterns of the MLCA estimates and the magnitudes of the MLCA estimates generally agree quite well with the H-W and traditional estimates for all three years. However, for the NLF category in 1995 and 1996, the traditional estimates of I are somewhat larger than either of the latent class model estimates. Further analysis suggests that this difference is due to a bias in the traditional estimation approach resulting from the failure of the parallel measures assumption.

U.S. Bureau of the Census (1985) shows that if the interview and reinterview processes have different reliabilities, then the traditional estimate of the index will be biased. For example, if the reliability of the reinterview

data is lower than the reliability of the interview data, the traditional test-retest reliability estimator will understate the actual reliability of the CPS data; *i.e.*, the CPS index of inconsistency will be too large.

Table 6
Comparison of MLCA, H-W, and Traditional Estimates of the Index of Inconsistency by Year and Labor Force Classification

Method of Estimation	Labor Force Classification			Aggregate Index
	Employed	Unemployed	Not in Labor Force	
1993				
Traditional estimation	8.16 (0.24)	33.49 (1.16)	9.96 (0.27)	11.05 (0.26)
H-W	7.37	34.93	10.07	10.78
MLCA	6.35	28.04	7.63	8.73
1995				
Traditional estimation	6.69 (0.44)	36.28 (2.85)	10.80 (0.56)	10.42 (0.53)
H-W	6.82	37	8.98	9.7
MLCA	6.06	36.19	7.2	8.72
1996				
Traditional estimation	5.93 (0.39)	35.97 (2.68)	11.95 (0.56)	10.61 (0.51)
H-W	5.67	39.46	7.55	8.56
MLCA	5.99	37.39	7.76	9.06

The CPS interview and reinterview will have different reliabilities if the error distributions for the two interviews are not equal. A test of this is possible by comparing the fit of a H-W type model with and without the restriction $\pi_{a|x} = \pi_{a'|x}$. The assumption of equal reliability is rejected if the difference between the likelihood ratio chi-squares for the two models exceeds a chi-square with 6 degrees of freedom. This test was rejected for 1995 and 1996 at the 10 percent level of significance. Thus, it appears that the difference in the NLF estimates for 1995 and 1996 may be due, in part, to bias in the traditional estimates of I.

Note further that the H-W and MLCA indexes agree quite well for 1995 and 1996, although they differ somewhat in 1993. However, as noted in the discussion of Table 5, the comparisons between the MLCA and H-W estimates for this year are confounded by the difference time periods used to construct the pre-1994 interview-reinterview data set. This could account for at least some of the discrepancy between the estimates for this year.

5. Summary and Conclusions

The primary goal of this research was to investigate the validity of MLCA estimates of CPS labor force classification error and to determine the efficacy of MLCA as an alternative to traditional methods for evaluating CPS data quality. We analyzed interview data from the CPS for the first quarter of three years – 1993, 1995, and 1996 – and conducted an additional analysis of the CPS unreconciled reinterview data for approximately the same time periods. The reinterview data provided another approach for estimating CPS classification error that, when compared with the MLCA estimates, helped to address the question of the validity of the MLCA approach.

Five dimensions of MLCA validity were addressed as follows:

1. **Model diagnostics.** We investigated a wide range of MLCA models with grouping variables defined by age, race, sex, education, mode of interview, and proxy/self response. The most parsimonious and best fitting model for all three years included one grouping variable defined by the proxy/self variable with four categories: all three waves conducted by self response, only two waves conducted by some self response, only two waves conducted by proxy response, and all three waves conducted by proxy response. For this class of models, the best model was Model 4 (see Table 1) which specified non-homogeneous and non-stationary transition probabilities and non-homogeneous response probabilities. This model provided an adequate fit to the data for all three years.

2. **Model Goodness of Fit Across Years of CPS.** Another indicator of model validity is its fit across independent samples of the same population. Assuming that labor force dynamics and the response probability structure for the CPS is stable across the span of four years, the same general model should fit all three years adequately. Model 4 displays multi-year goodness of fit (see Table 1). In addition, other grouping variables were tested in the study, yet the proxy/self variable model emerged as the best variable for all three years.

3. **Agreement Between the Model and Test-Retest Estimates of Response Probabilties.** Using the unreconciled interview-reinterview data from the CPS for the time periods pre-1994, 1995, and 1996, we applied the H-W method to estimate the response probabilities and compared these with the MLCA estimates. There was good agreement for 1995 and 1996, the two years for which the time periods for the reinterview data and the CPS data were closely matched (see Table 5). For 1993, we observed small but significant differences between MLCA estimates and the corresponding H-W estimates. These differences might be explained by differences in the time periods, since the reinterview data predated the CPS interview data by some years.

4. **Agreement Between the Model and Test-Retest Estimates of Inconsistency.** We compared MLCA model-based estimates of the index of inconsistency with the corresponding direct estimates from the CPS

reinterview program. The two sets of estimates agree fairly well for all three years, with the exception of the NLF category (see Table 6). For 1995 and 1996, the differences can be partly explained by the bias in the traditional estimator resulting from the failure of the parallel measures assumption. The H-W method, which does not require the assumption of parallel measures, produces estimates of the index that agree well with MLCA estimates for 1995 and 1996. For 1993, the difference between MLCA and H-W estimates may be due to the difference in the time periods for the reinterview and the CPS data sets.

5. **Plausibility of the Patterns of Classification Error.** The MLCA estimates of misclassification probabilities appear to be plausible. The estimates across proxy/self groups were consistent with prior expectations that lower error rates should be observed for self respondents than for proxy respondents. In addition, the largest error rates were observed for the unemployed population and the magnitudes of these estimates were consistent with those of previous studies – for *e.g.*, Fuller and Chua 1985; Abowd and Zellner 1985; Porterba and Summers 1986; and Sinclair and Gastwirth 1996 (see Table 3).

In Summary, we found no evidence from these analyses to question the validity of the MLCA approach. The method performed well in all five validity tests. We therefore recommend that the MLCA method be considered as an alternative method for evaluating the accuracy of the CPS labor force estimates. The strong agreement between the MLCA and H-W estimates supports the validity of the H-W method as well. We recommend that both methodologies be considered in future studies of CPS data quality.

Although the MLCA approach performed well in our tests, we recommend caution in applying the methodology in other settings. In our analysis, reinterview data provided a means for assessing the validity of the MLCA estimates. However, reinterview data are typically not available in panel surveys and, consequently, analysts may only be able to apply criteria (1), (2), and (5) above to check model validity. The Markov assumption is key to the MLCA approach. Some panel data may seriously violate this assumption. Fortunately, failure of Markov assumption appears not to be an important factor in the validity of MLCA estimates of CPS labor force classification error (*cf.* Table 4).

References

Abowd, J., and Zellner, A. (1985). Estimating gross labor-force flows. *Journal of Business and Economic Statistics*, 3, 254-283.

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.

Akerlof, G.A., and Main, G.M. (1980). Unemployment spells and unemployment experience. *The American Economic Review*, 70, 3, 885-893.

Bailar, B.A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

Biemer, P., Bushery, J. and Flanagan, P. (1997). An Application of Latent Markov Models to the CPS. Internal U.S. Bureau of the Census Technical Report.

Biemer, P., and Forsman, G. (1992). On the quality of reinterview data with applications to the Current Population Survey. *Journal of the American Statistical Association*, 87, 420, 915-923.

Bohrnstedt, G.W. (1983). Measurement. *Handbook of Survey Research*, (Eds., P.H.l Rossi, R.A. Wright and A.B. Anderson). New York: Academic Press.

Bushery, J., and Kindelberger, K. (1999). Simulation Examples for MLCA Analysis. Internal U.S. Bureau of the Census Memorandum, Washington, DC, 70-122.

Chua, T.C., and Fuller, W.A. (1987). A model for multinomial response error appliced to labor flows. *Journal of the American Statistical Association*, 82, 397, 46-51.

Clogg, C., and Eliason, S. (1985). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16, 8-14.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 210, 37-46.

Corak, M. (1993). Is unemployment insurance addictive? Evidence from the benefit durations of repeat users. *Industrial and Labor Relations Review*, 47, 1, 67-72.

Forsman, G., and Schreiner, I. (1991). The design and analysis of reinterview: An overview. *Measurement Errors in Surveys*, (Eds., P.P. Biemer*, et al.*). New York: John Wiley & Sons, Inc., 279-302.

Fuller, W., and Chua, T.C. (1985). Gross change estimation in the presence of response error. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*. Washington, D.C., U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, 65-77.

Heckman, J.J., and Borjas, G.J. (1980). Does unemployment cause future unemployment? Definitions, questions, and answers from a continuous time model of heterogeneous and state dependence. *Economica*, 47, 247-283.

Hess, J., Singer, E. and Bushery, J. (2000). Predicting test-retest reliability from behavior coding. *International Journal of Public Opinion Research*, II, 4, 346-360.

Hui, S.L., and Walter, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, 167-171.

Kleinbaum, D.G., Kupper, L.L. and Muller, K.E. (1988). *Applied Regression Analysis and Other Multivariate Methods*. Boston: PWS-KENT Publishing Co.

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.

Lynch, L.M. (1989). The youth labor market in the eighties: determinants of re-employment probabilities for young men and women. *The Review of Economics and Statistics*, 37-45.

Meyers, B.D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 3, 385-390.

Moore, J.C. (1988). Self/proxy response status and survey response quality. *Journal of Official Statistics*, 4, 2, 122-155.

O'Muircheartaigh, C. (1991). Simple response Variance: Estimation and Determinants. *Measurement Errors in Surveys*, (Eds., P. Biemer *et al*.). New York: John Wiley & Sons, Inc., 551-574.

Poterba, J., and Summers, L. (1986). Reporting errors and labor market dynamics. *Econometrics*, 54, 6, 1319-1338.

Poterba, J., and Summers, L. (1995). Unemployment benefits and labor market trasitions: A multinomial logit model with errors in classification. *The Review of Economics and Statistics*, 77, 207-216.

Poulsen, C.S. (1982). Latent Structure Analysis with Choice Modeling Applications. Doctoral dissertation, Wharton School, University of Pennsylvania.

Rothgeb, J. (1994). Revisions to the CPS Questionnaire: Effects on Data Quality, U.S. Bureau of the Census. CPS Overlap Analysis Team Technical Report 2, April 6.

Schreiner, I. (1980). Reinterview Results from the CPS Independent Reconciliation Experiment (Second Quarter 1978 through Third Quarter 1979). Internal U.S. Bureau of the Census Report.

Shockey, J. (1988). Adjusting for response error in panel surveys, a latent class approach. *Sociological Methods and Research*, 17, 1, 65-92.

Sinclair, M., and Gastwirth, J. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: Application to labor force data. *Journal of the American Statistical Association*, 91, 961-969.

Sinclair, M., and Gastwirth, J. (1998). Estimates of the errors in classification in the labour force survey and their effects on the reported unemployment rate. *Survey Methodology*, 24, 2, 157-169.

Singh, A.C., and Rao, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian labour force survey. *Journal of the American Statistical Association*, 90, 430, 478-488.

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

U.S. Bureau of the Census (2000). Current Population Survey: Design and Methodology. U.S. Bureau of the Census Technical Paper 63, Washington, D.C.: Government Printing Office.

Van de Pol, F., and De Leeuw, J. (1986). A latent markov model to correct for measurement error. Sociological Methods and Research, 15, 1-2, 118-141.

Van de Pol, F., and Langeheine, R. (1997). Separating change and measurement error in panel surveys with an application to labor market data. *Survey Measurement and Process Quality*, (Eds., L. Lyberg, *et al.*). New York: John Wiley & Sons, Inc.

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

Wiggins, L.M. (1973). *Panel Analysis, Latent Probability Models for Attitude and Behavior Processing*, Amsterdam: Elsevier S.P.C.