Survey Methodology

December 2004
How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line 1 800 263-1136
National telecommunications device for the hearing impaired 1 800 363-7629
Depository Services Program inquiries 1 800 700-1033
Fax line for Depository Services Program 1 800 889-9734
E-mail inquiries infostats@statcan.ca
Website www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.
Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.
An Analysis of Classification Error for the Revised Current Population Survey Employment Questions

Paul P. Biemer

Abstract

The reduced accuracy of the revised classification of unemployed persons in the Current Population Survey (CPS) was documented in Biemer and Bushery (2000). In this paper, we provide additional evidence of this anomaly and attempt to trace the source of the error through extended analysis of the CPS data before and after the redesign. The paper presents a novel approach decomposing the error in a complex classification process, such as the CPS labor force status classification, using Markov Latent Class Analysis (MLCA). To identify the cause of the apparent reduction in unemployed classification accuracy, we identify the key question components that determine the classifications and estimate the contribution of each of these question components to the total error in the classification process. This work provides guidance for further investigation into the root causes of the errors in the collection of labor force data in the CPS possibly through cognitive laboratory and/or field experiments.

Key Words: Survey redesign; Measurement error; Latent class analysis; Unemployment rate; Specification error.

1. Introduction

The Current Population Survey (CPS) is a monthly survey of approximately 60,000 households conducted by the U.S. Bureau of the Census for the Bureau of Labor Statistics (BLS). The primary purpose of the survey is to provide estimates of employment, unemployment, and other characteristics of the general U.S. labor force population. Estimates of the size, composition, and dynamic characteristics of the labor force are published each month by BLS and comprise one of the Nation’s key economic indicators.

In January 1994, a revised questionnaire was introduced in the CPS to address the recommendations by the Levitan Commission in the late 1970s to convert the mode of interview for the CPS from paper and pencil questionnaire to computer-assisted interviewing methods, to clarify some of the questions on employment, as well as for a number of other reasons described in Rothgeb (1994). The overall objective of the redesign was to improve the quality of the data collected in the CPS. The CPS questionnaire had remained essentially unchanged since the last major revision in 1967.

The revised CPS questionnaire was introduced after considerable research and testing that began in the mid-1980s. The purpose of the testing was to evaluate the quality and operational feasibility of various redesign options including moving the CPS from a paper and pencil questionnaire format to computer assisted interviewing. During these years of testing, more than 100,000 persons were interviewed in the various studies that were conducted (Rothgeb 1994). The CPS redesign research program culminated in a large national study (referred to in the literature as the CATI/CAPI Overlap or CCO Field Test) that was conducted in 1993. The key component of this test consisted of a computer assisted survey of approximately 12,000 households implementing revised CPS interviewing procedures and the revised questionnaire. This survey, referred to in this report as the Parallel Survey, was conducted from July 1992 to December 1993 concurrently with the ongoing CPS survey which used the original questionnaire. This type of split panel design makes it possible to estimate the effect of the redesign changes on the CPS labor force estimates.

A number of papers and reports were published documenting the findings from the CCO Field Test (Cohany, Polivka and Rothgeb 1994; Rothgeb 1994; Polivka 1994; Kostanich and Cahoon 1994; Miller 1994; Thompson 1994; Dippo, Polivka, Creighton, Kostanich and Rothgeb 1994). One key finding from this research was that the Parallel Survey unemployment rate and the labor force participation rate were higher than in the CPS. The higher unemployment and labor force participation rates associated with the revised questionnaire were explained primarily by changes in the definition of employment. The revised questionnaire has a broader approach to both work and job search activities, which would tend to classify more persons as “in the labor force” and, thus, more persons who are not working as unemployed rather than out of the labor force (see, for example, Polivka 1994 and Rothgeb 1994).

The increase in the unemployment rate due to the new design was originally estimated at about one-half percentage point. However, further analysis of the Parallel Survey data...
called that estimate into question and subsequently a report was release estimating the increase to be less than one-tenth percentage point (Polivka and Miller 1994). The concerns raised in the subsequent reports regarding the utility of the Parallel Survey data for assessing the effect of the redesign are discussed further below and will be considered in our analysis of these data.

An independent analysis conducted by Biemer and Bushery (2000) revealed an anomaly in the revised CPS labor force data that had not been detected by any of the previous research on the CPS redesign. Using a Markov latent class analysis (MLCA) approach, Biemer and Bushery compared the accuracy of labor force classifications under the original and revised designs by estimating and comparing the error rates using the 1993 CPS data and the 1995 and 1996 CPS data. They defined labor force classification accuracy as the probability that a person who is truly in some labor force category, say category $a$, is classified as being in $a$ by the CPS; i.e., $P_r(\text{classified in } a \mid \text{true in } a)$. For example, the classification accuracy for unemployment is the probability a person who is truly unemployed, according to the CPS definition, is correctly classified as unemployed by the CPS classification rules.

In Table 2 of their paper, Biemer and Bushery report that the classification accuracy for unemployment dropped by 5.7 percentage points, from approximately 81.8 percent (s.e. = 0.90) in 1993 to 76.1 (s.e. = 1.2) in 1995 and 74.4 percent (s.e. = 1.2) in 1996. These results suggest that the redesigned CPS misclassifies the true unemployed at a higher rate than the old CPS design. The authors first considered that this result could be an artifact of the MLCA methodology. As shown below, MLCA does not require a true or “gold standard” measurement of employment to estimate classification error. Rather the method relies a model describing the true month to month changes in employment status and as well as for the process of classifying individuals into labor force categories. It is possible that labor force transitions that deviate from the model specification could be regarded as misclassifications in the estimation process.

To check the validity of the MLCA results, the authors conducted a series of analyses using traditional estimation approaches, analysis of the error by population groups, comparisons of the error estimates to other published estimates, and simulations to assess the effect of model failure on the results. As an example, there is evidence that the test-retest reliability of the unemployment category decreased after the redesign. Prior to the redesign, the index of inconsistency (The index of inconsistency is a measure of unreliability traditionally used at the Census Bureau. It is equal to $1 - \kappa$ where $\kappa$ is Cohen’s kappa coefficient (Cohen 1960) for the unemployed labor category averaged 30 percent for the period 1992–1993. Following the redesign, the index of inconsistency increased to almost 40 percent for the period 1995–1996. These analyses support their claim that the accuracy of the CPS methodology for classifying unemployed persons declined after the redesign.

In their discussion of the results, the authors speculated that the drop in classification accuracy could indicate a problem with the revised unemployment questions. That is, the revised unemployment questions may be subject to greater classification error and, thus, less classification accuracy. Another possibility they considered is change in the characteristics of the unemployed populations from 1993 to 1995 and 1996. Since the unemployment rate dropped from 1993 to 1996, it is possible that persons who would be more accurately classified by the CPS system left the ranks of the unemployed, leaving persons who would be less accurately classified in the category. This hypothesis could be tested by estimating the accuracy rates for the two methodologies for the same time period. The Parallel Survey offers a means to conduct such an analysis.

The current paper continues the investigation of the reduction in MLCA unemployment classification accuracy rates observed by Biemer and Bushery. The current analysis uses MLCA models very similar to those used by Biemer and Bushery for estimating the classification accuracy for the original and revised versions of the CPS questionnaire. However, the time period considered here is expanded to include the 15 months prior to and following the introduction of the revised questionnaire: a total of 30 contiguous months. In addition, data from the Parallel Survey from the period January 1993 through December 1993 is used to compare the employment accuracy for original and revised questionnaire for the same time period.

Our analysis focuses on a labor force classification variable that is derived from a number of questions on the employment section of the CPS questionnaire. This variable is often referred to as a “recoded” labor force variable since it is determined by mapping a pattern of CPS responses to questions about employment onto particular labor force categories such as employed – at work, employed – not at work, unemployed – looking for work, and so on. Biemer and Bushery used a three-category employment classification variable: employed (EMP), unemployed (UEM), and not in the labor force (NLF). For the present analysis, a four-category variable is used that subdivides the UEM category into unemployed-on layoff (UEM-LAYOFF) and unemployed-looking for work (UEM-LOOKING). This is done as a first step toward isolating the source of the apparent inaccuracy in unemployment classification. However, further decomposition of these categories will be necessary to arrive at the root source of the error as will be shown subsequently.
In section 2 we describe the CPS labor force concepts that are most relevant to our study and the structure of the data sets in the analysis. In section 3 we review the MLCA estimation methodology and models used by Biemer and Bushery in their analysis and describe the application of their methodology for the present purposes. In section 4 we present the results of our analysis and what they suggest regarding the source of the classification error in the new questionnaire. Finally, section 5 provides a summary of the key findings and our conclusions from the study.

2. Data and Concepts

2.1 The Data Sets for Our Study

Except for the Parallel Survey, the CPS data in our analysis were downloaded from the National Bureau of Economic Research (NBER) website (www.nber.org). This website contains microdata for the CPS for every month from January 1976 through December 2004. The MLCA approach was applied directly to these microdata without the need for supplementary data or data external to the CPS.

In the preliminary analysis, we investigated the CPS classification accuracy for a six-year period: January 1992 through December 1997. That analysis was aimed at determining whether the anomaly first noted in Biemer and Bushery (2000) is a transient phenomenon affecting only the months immediately following the introduction of the new questionnaire or whether it persisted for some years after the new questionnaire was introduced. If temporary or transient, the anomaly might be related to problems during the phase-in of the new design; for example, interviewer training or issues related to the startup of data collection. However, evidence of a persistent, continuing effect could suggest problems with the survey design; for example, the questionnaire, interviewing procedures, or the recoding algorithm.

By applying MLCA across all months from 1992 through 1997 we determined that, although the magnitude of the reduction in accuracy varies somewhat from month to month, it does indeed persist for all months following the introduction of the revised questionnaire. The results confirmed Biemer and Bushery’s conjecture of a systemic effect possibly linked to the new unemployment questions introduced in January 1994.

Due to space considerations, in this paper we present results from a somewhat shorter time frame than considered in the preliminary analysis, viz., the years 1992, 1993, 1994, and 1995. This time period covers two years of the CPS using the original questionnaire and two years using the revised questionnaire. In addition, we will also present some results from an MLCA of the 1993 Parallel Survey data that can be compared with results from the main CPS.

The data sets in our study are quite large. Each estimate of classification error we obtain is based upon all households that were interviewed in the CPS for three consecutive months. Across the four years in our analysis, the total number of households responding for all three months in any three-month period varies from about 37,000 to more than 40,000. For the 1993 Parallel Survey, the number of households satisfying this criterion is approximately 10,000. The estimates we produce are appropriately weighted for probabilities of selection and other post-survey adjustments and, therefore, reflect the response probabilities of the published CPS estimates. Weights were constructed by taking an average weight across the three consecutive months that were combined to form a longitudinal record for the analysis (unweighted analyses were also conducted and the results were very similar to the weighted analysis. This suggests the choice of weights has little effect on the study outcomes).

Because of a problem in the identification variables required for linking households for the months June 1995 through December 1995, it was not possible to include these months in our analysis. Further, since our conclusions would not change by including data from the 1996 or later years of the CPS, we confine our analysis to 15 months prior and 15 months following the introduction of the revised questionnaire. Thus, for most of the analysis to follow, we will provide averages of estimates from August 1992 through December 1993 for the original questionnaire and from January 1994 through May 1995 for the revised questionnaire (note that since our estimates are based upon a moving average of three consecutive months, seasonal variations in the labor rates and transitions probabilities are accounted for in the estimates of classification error).

2.2 Labor Force Concepts

The revised CPS questionnaire was introduced in 1994 to improve the overall quality of labor market information through extensive question changes and through the use of computer technology in the data collection. In the following, we describe a few concepts that were affected by the questionnaire redesign and that are relevant for the current analysis.

**Employed.** The labor force questions in the original questionnaire began with the question “What were you doing most of LAST WEEK (working, keeping house, going to school, or something else)?” Interviewers were allowed to modify the parenthetical part of this question according to the age of the respondent. In some cases, the word “work” or “working” was not part of the question. As an example, if the respondent looked of student-age, the interviewer was allowed to leave out the word “working.” The revised questionnaire replaced this question with two
questions: “Does anyone in this household have a business or a farm?” and “LAST WEEK, did you do ANY work for (either) pay (or profit)?” where the parenthetical parts of the question are read if anyone in the response to the first question is “yes.” Further, additional questions were added to clarify whether earnings or profits were received from the family business or farm. Thus, the revised questionnaire concept of employment appears to be somewhat broader and better defined than the original questionnaire concept.

**Unemployed.** The definition of unemployment was slightly modified in the revised questionnaire. In the original questionnaire, persons waiting for a new job to start were classified as unemployed. Under the revised questionnaire definition, a person is unemployed only if all of the following are true: (1) without a job, (2) actively seeking work or on layoff from a job and expecting recall within the next six months, and (3) currently available to take a job (except for a possible temporary illness).

**On Layoff.** Persons on layoff are defined as persons separated from a job and who are awaiting a recall to return to that job. The original questionnaire did not consider or collect information on the expectation of recall. This was problematic because to most people, the term “layoff” could mean permanent termination from the job rather than the temporary loss of work economists are trying to measure.

**Job Search Methods.** To be counted as unemployed and looking for work, a person must have engaged in an active job search during the four weeks prior to the survey. The revised questionnaire includes a somewhat broader question about job search methods with expanded and restructured response categories to allow interviewers to more easily record and distinguish between active and passive job search activities. In addition, it provides additional followup questions for those who respond “nothing” or “don’t know.”

**Reference Week.** While the original questionnaire referred to LAST WEEK, the reference period was never explicitly defined. The revised questionnaire provides specific dates of the reference week.

We will refer to these changes later in the report when we discuss the differences in the classification error and specification error between the revised and original questionnaires.

As previously noted, Biemer and Bushery focused on a three-category labor force recoded variable with categories: employed (EMP), unemployed (UEM), and not in the labor force (NLF). For the present analysis, we used an expanded recoded variable also available on the CPS public use data files. This variable divides the UEM category into two categories corresponding to persons on layoff (LAYOFF) and persons looking for work (LOOKING). The seven-category variable also divides the EMP and NLF categories into subcategories; however, this level of detail in the EMP and NLF categories is not needed in our analysis. Thus, the seven-category variable will be collapsed to a four-category variable corresponding to EMP, UEM-LOOKING, UEM-LAYOFF, and NLF. The correspondence between the three- and four-category variables is shown in Figure 1.

### 3. Latent Class Models for CPS Classification Error

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 describe the basic model proposed by Biemer and Bushery (2000) and its extensions for application in the current analysis.

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
The $i^{th}$ population member is in group 1, $G_i = 2$ for group 2 and so on. Let $X_{gi}, Y_{gi}$, and $Z_{gi}$ denote the true labor force classifications for the $i^{th}$ person in group $G = g$ (for $g = 1, \ldots, L$ and $i = 1, \ldots, n_g$) where $X_{gi}$ is defined as

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

with analogous definitions for $Y_{gi}$ and $Z_{gi}$ for periods 2 and 3 respectively. Consistent with the conventions of the LCA literature, we will drop the subscripts from the variables to simplify the notation.

Let $\pi_{x,y,z|g}$ denote $\Pr(X = x, Y = y, Z = z \mid G = g)$, let $\pi_{x,y|g}$ denote $\Pr(Y = y \mid X = x, G = g)$ and let $\pi_{z|g,y,x}$ denote $\Pr(Z = z \mid Y = y, X = x, G = g)$. 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 $\pi_{x,y,z|g}$ which may be written as

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

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

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

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.

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

$$A_{gi} = \begin{cases} 
1 & \text{if person } (g, i) \text{ is classified as EMP in time period 1} \\
2 & \text{if person } (g, i) \text{ is classified as UEM – LAYOFF in time period 1} \\
3 & \text{if person } (g, i) \text{ is classified as UEM – LOOKING in time period 1} \\
4 & \text{if person } (g, i) \text{ is classified as NLF in time period 1}
\end{cases}$$

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 $\pi_{a|g} = \Pr(A = a \mid X = x)$, with analogous definitions for $\pi_{b|g}$ and $\pi_{c|g}$. Thus, $\pi_{a|g} = \pi_{a|g}$ is the probability that the CPS classifies a person in group $g$ as employed ($A = 1$) when the true status is unemployed – on layoff ($X = 2$). Likewise, $\pi_{a=2,g=x=2}$ is the probability that the CPS correctly classifies a person in group $g$ as unemployed – on layoff.

Finally, we assume

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

or that classification error in the observed labor force status is independent across the three months.

The CPS labor force classifications for each month of a three consecutive month interval are the outcome variables in our analysis. Let $A, B, \text{ and } C$ denote the observed classifications and let $X, Y, \text{ and } Z$ denote the (unobserved) true classifications for Month 1, Month 2, and Month 3, 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:

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

Extensions to more than one grouping variable are straightforward.

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

$$\Pr(GABC) = C \prod_{g,a,b,c} \pi_{g,a,b,c} \quad (5)$$

where $C$ is the multinomial constant and $\Pi$ denotes the product of the terms over the subscripts $g, a, b, \text{ and } c$. Under the assumptions made previously, the model parameters are estimable using maximum likelihood estimation methods. Van de Pol and de Leeuw (1986) provide the formula for applying the E-M algorithm to estimate the parameters of this model and describe the conditions for their estimability. The E-M software (Vermunt 1997) was used to fit the MLCA models.

In their investigation of the validity of MLCA estimates for analyzing CPS labor force classification error, Biemer and Bushery analyzed CPS data collect during the first quarter of each of three years – 1993, 1995, and 1996. They also conducted several types of analysis using the CPS un-reconciled reinterview data for the same time period. The reinterview analysis provided another approach for
estimating CPS classification error as well as evidence of the validity of the MLCA approach. Their evaluation of MLCA validity considered five criteria: (1) model diagnostics, (2) model goodness of fit across years of CPS, (3) agreement between the model and test-retest estimates of response probabilities, (4) agreement between the model and test-retest estimates of inconsistency, and (5) plausibility of the patterns of classification error. The MLCA method performed well in all five tests. For example, the same model provided the best fit of the data for each year analyzed, there was good agreement between the latent class estimates of reliability and those derived from traditional test-retest methodology; and the estimated error rates were consistent with those of previous studies – for e.g., Chua and Fuller 1987; Abowd and Zellner 1985; Portera and Summers 1995; and Sinclair and Gastwirth 1998.

Ostensibly, the Markov assumption seems very unlikely to hold for labor force data. As an example, persons who are unemployed in months 1 and 2 of a consecutive three-month period may not have the same probability of being unemployed in a month 3 as persons who just became unemployed in month 2. The former group could contain more chronically unemployed persons than the group entering unemployment in month 2. Further, the group just entering unemployment in month 2 could contain a higher proportion of people temporarily out of work while changing jobs. Biemer and Bushery considered the consequences for the MLCA estimates of misclassification when the Markov assumption is violated.

Using simulation, Biemer and Bushery found that the bias in MLCA estimates of classification probabilities depends upon the severity of the departures of the CPS data from the Markov assumption. They defined two parameters, \( \lambda_1 \) and \( \lambda_2 \), which are ratios of conditional probabilities. \( \lambda_1 \) is the ratio of the probability of being employed in period 3 for a person with an (EMP, UEM) pattern for periods 1 and 2, respectively, divided by the probability of being employed in period 3 for a person with a (EMP, EMP) pattern. Similarly, \( \lambda_2 \) is the ratio of the probability of being employed in period 3 for a person with an (UEM, UEM) pattern to the probability of being employed in period 3 for a person with a (EMP, UEM) pattern. Note that when \( \lambda_1 = \lambda_2 = 1 \), the Markov assumption holds exactly and greater departures of \( \lambda_1 \) and \( \lambda_2 \) from 1 correspond to greater departures of the data from the Markov assumption. Biemer and Bushery found that over a fairly wide range of values for \( \lambda_1 \) and \( \lambda_2 \), the absolute bias in the MLCA estimates of unemployment classification accuracy never exceeded 3 percentage points. For example, in the extreme case of a Markov assumption violation, the expected value of an MLCA estimate of unemployment accuracy would be 77 percent when the true parameter value is 80 percent. Their results suggest that, for the CPS application, MLCA is fairly robust to failures of the Markov assumption to hold.

Although it is virtually impossible to prove their validity, MLCA error estimates can be quite useful for identifying survey questions that are prone to classification error; i.e., flawed questions. For example, Biemer (2004) and Biemer and Wiesen (2002) demonstrate the utility of MLCA methodology for identifying question problems and classification process deficiencies in large scale surveys. Notwithstanding that the MLCA assumptions may be violated to an unknown extent, its usefulness as a tool for exploring a number of important questionnaire design issues has been well-documented. For the present application, MLCA will be used to develop and test hypotheses regarding the sources of the anomaly reported by Biemer and Bushery for 1994 CPS redesign.

The MLCA model used in the present analysis is essentially the same model selected by Biemer and Bushery for their analysis. To account for population heterogeneity, they considered a number of demographic and other explanatory variables that might be highly correlated with classification error. The best performing variable a proxy or self-response indicator variable denoted by \( P \) where

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

Their empirical findings showed this variable to be strongly related not only to reporting accuracy, but also current employment status and month to month employment transitions. For example, responses for the PROXY group were considerably less accurate than for the SELF group and, further, the PROXY group had somewhat higher unemployment than the SELF group.

The MLCA model also allows transition probabilities to vary by \( P \) (referred to as group heterogeneity) as well as by time periods (referred to as non-stationary transitions). In addition, the model assumes that response probabilities \( \pi_{apx}, \pi_{byp}, \) and \( \pi_{cyp} \) are group-heterogeneous but are equal for all three months in the time interval. This leads to the following model for describing the cell probabilities in the \( PABC \) table:

\[
\pi_{p,a,b,c} = \sum_{x,y,z} \pi_{p} \pi_{x|p} \pi_{y|x} \pi_{z|y} \pi_{a|zy} \pi_{b|yp} \pi_{c|yp} \pi_{x|yp} \pi_{y|yp} \pi_{z|yp} \pi_{A|PX} \pi_{B|PY} \pi_{C|PY} \tag{6}
\]
where \( \pi_{APX}^{4X} = \Pr(A = b \mid P = p, X = y) \) with similar definitions for \( \pi_{APX}^{2X} \) and \( \pi_{APX}^{1X} \). That is, the three sets of response probabilities are equal to \( \pi_{APX}^{4X} \).

Note that for the present analysis, interest is focused on the overall response probabilities associated with the revised and original questionnaires and not the variation in error rates across proxy groups. Therefore, our analysis focuses on the overall accuracy of response, i.e., \( \pi_{APX}^{4X} \) or the mean response probability for the four levels of \( P \) combined.

4. Comparison of Revised and Original Questionnaire Classification Error Probabilities

4.1 Reduction in UEM Classification Accuracy for the Revised Questionnaire

As mentioned in section 2, the CPS data sets for this analysis are monthly samples from August 1992 through May 1995. Figure 2 shows how this time interval was divided into 30 overlapping three-month intervals: 15 for the original questionnaire and 15 for the revised questionnaire. The intervals are numbered in the table for later reference. For example, time interval 1 covers the period from August 1992 through October 1992 in which the original questionnaire was in use. Therefore, this time interval can provide one estimate of the response probabilities, \( \pi_{APX}^{4X} \), for the model in (6). Since there are 30 time intervals across the entire 34-month period in our analysis, 30 estimates of \( \pi_{APX}^{4X} \) can be formed from these consecutive overlapping time intervals: 15 estimates for the original questionnaire and 15 estimates for the revised questionnaire.

To obtain a more stable estimate of \( \pi_{APX}^{4X} \) for each questionnaire, the 15 estimates corresponding to the 15 time periods per questionnaire in Figure 2 were averaged. These estimates are shown in Tables 1 and 2. Since they are based on simple random sampling assumptions, the standard errors in the tables do not account for the unequal weighting and clustering effects of the CPS. Since the average CPS design effect is about 1.5 for estimates of unemployment, the standard errors in the tables are probably understated by 20 percent or less. This level of bias in the standard errors is inconsequential for the purposes of this paper due to the extremely large sample sizes in the analysis.

Table 1 compares the MLCA estimates of the classification error probabilities for the original and revised questionnaire versions for the three-category labor force classification scheme used by Biemer and Bushery. The first column of the table is the true (or latent) category, the second column is the observed (or CPS) category, and the cell entries are the response probabilities estimated from the MLCA using model (6). For each true class (EMP, UEM, or NLF), the accuracy rate is the cell corresponding to the observed category with the same label. For example, the accuracy of classifying persons who are truly employed is 98.68 percent (for the original questionnaire) and 98.84 percent for the revised questionnaire. Note that this entry corresponds to the cell where both the true category and the observed category are EMP. The other cells for EMP in column 1 are the error rates for EMP. For example, the MLCA estimate of the probability CPS classifies a person as UEM who is truly EMP is 0.42 for the original questionnaire and 0.39 for the revised questionnaire. The other cell entries are interpreted analogously.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Interval</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>…†</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>1 (Old), 16 (New)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>…†</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2 (Old), 17 (New)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>…†</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3 (Old), 18 (New)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>…†</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4 (Old), 19 (New)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>…†</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>…†</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>…†</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>13 (Old), 28 (New)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>…†</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>14 (Old), 29 (New)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>…†</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>15 (Old), 30 (New)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>…†</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

† The “…” symbol is used in this table to indicate that the pattern established for the preceding months continues for the remaining months.

Figure 2. The 30 Three-Month Time Intervals Analyzed for the Revised and Original Questionnaires.
Table 1
Comparison of CPS Labor Force Response Probabilities for the Original and Revised Questionnaires

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EMP</td>
<td>EMP</td>
<td>98.68</td>
<td>98.84</td>
<td>–0.15</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>UEM</td>
<td>0.42</td>
<td>0.39</td>
<td>0.03</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>NLF</td>
<td>0.90</td>
<td>0.78</td>
<td>0.13</td>
<td>0.16</td>
</tr>
<tr>
<td>UEM</td>
<td>EMP</td>
<td>8.23</td>
<td>10.57</td>
<td>–2.34*</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>UEM</td>
<td>79.06</td>
<td>73.50</td>
<td>5.56*</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>NLF</td>
<td>12.71</td>
<td>15.93</td>
<td>–3.22*</td>
<td>0.26</td>
</tr>
<tr>
<td>NLF</td>
<td>EMP</td>
<td>2.14</td>
<td>1.99</td>
<td>0.15</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>UEM</td>
<td>1.43</td>
<td>1.56</td>
<td>–0.13</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>NLF</td>
<td>96.43</td>
<td>96.45</td>
<td>–0.02</td>
<td>0.18</td>
</tr>
</tbody>
</table>

* Significant at \( \alpha = 0.001 \).

Table 2
Comparison of Two Unemployed Subcategories for the Original and Revised Questionnaires

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>UEM–LAYOFF</td>
<td>EMP</td>
<td>16.32</td>
<td>26.67</td>
<td>–10.35*</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>UEM–Layoff</td>
<td>61.30</td>
<td>55.63</td>
<td>5.66*</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>UEM–Looking</td>
<td>17.61</td>
<td>8.41</td>
<td>9.20*</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>NLF</td>
<td>4.77</td>
<td>9.29</td>
<td>–4.52*</td>
<td>0.28</td>
</tr>
<tr>
<td>UEM–LOOKING</td>
<td>EMP</td>
<td>7.03</td>
<td>7.51</td>
<td>–0.48</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>UEM–Layoff</td>
<td>1.43</td>
<td>0.65</td>
<td>0.38</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>UEM–Looking</td>
<td>78.00</td>
<td>74.61</td>
<td>3.39*</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>NLF</td>
<td>13.94</td>
<td>17.23</td>
<td>–3.29*</td>
<td>0.18</td>
</tr>
</tbody>
</table>

* Significant at \( \alpha = 0.001 \).

Table 2 shows the same set of estimates for the truly employed population only in somewhat greater detail. In this table, we considered the two primary subclassifications of unemployed: UEM-LAYOFF and UEM-LOOKING. This table provides information regarding the source difference in accuracy rates between the two questionnaire versions. We first consider the misclassification of true LAYOFF persons (top half of the table) and then consider the LOOKING persons (bottom half of the table).

For persons on layoff, classification accuracy appears to have dropped an average of 5.66 percentage points with the introduction of the revised questionnaire: from 61.30 percent to 55.63 percent. However, the patterns of classification error also changed. For the original questionnaire, the probability that a person on layoff is misclassified as looking for work is estimated at about 18 percent. The corresponding estimate for the revised questionnaire is less than half that: 8.5 percent. In addition, the data suggests that misclassification of unemployed persons on layoff as either employed or not in the labor force increased by 10.35 and 4.52 percentage points, respectively.

Now consider persons who are truly looking for work in the bottom half of Table 2. According to the MLCA model, classification accuracy for the redesigned CPS decreased significantly from 78.00 to 74.61 percent. Most of the misclassification is attributed to misclassifying persons looking for work as NLF. This result would arise, for example, if the questions regarding active and passive job search activities are prone to error. To further investigate this finding, we conducted an analysis of each of the questions used to determined the LOOKING recode. In the next section, we first consider the sources of error in the LAYOFF classification and then investigate the sources of error for the LOOKING classification.

4.2 Specific Questions Responsible for the Reduction in LAYOFF Accuracy

4.2.1 Decomposition of the LAYOFF Recode

Individuals in the CPS are classified as LAYOFF on the basis of their responses to five questions in the original questionnaire and eight questions in the revised questionnaire. These questions are listed in Figure 3. Initially, we consider which questions or combinations of questions contribute most to the error rate observed in Table 2 for the LAYOFF recoded variable and then show how MLCA models can be applied to estimate the contributions to classification error of individual questions that are used to classify an individual as LAYOFF. The methodology employed for this is similar to the MLCA approach used previously for estimating the aggregate classification error. We will describe this technique in terms of the LAYOFF classification, but it will be applied subsequently to
decompose the error in both the LAYOFF and LOOKING classification processes.

First, we combine the questions in Figure 3 using the logical operators such as “and,” “or,” “if-then-else,” etc. to form a set of dichotomous “compound” questions with the property that each compound question must be answered positively in order for an individual to be classified as LAYOFF by the CPS classification process. Let \( Q_k \) denote the outcomes to the \( K \) compound questions that were formed for the LAYOFF classification, where \( Q_k = 1 \) denotes a positive outcome and \( Q_k = 2 \) denotes a negative outcome. Then an individual in the CPS is classified as LAYOFF if and only if \( Q_k = 1 \) for \( k = 1, \ldots, K \).

For each classification, \( Q_k \) there is a corresponding true, unobservable (latent) classification, \( T_k \) defined in analogy to \( Q_k \); i.e., an individual is truly on layoff by the CPS definition if and only if \( T_k = 1 \) for \( k = 1, \ldots, K \). Next, we will use MLCA to estimate the misclassification error rates for each compound question \( Q_k \) by treating these as indicators for the unknown true latent characteristics, \( T_k \).

The probability of an error in the classification of LAYOFF can be written as

\[
Pr(Q_k = 2 \mid T_k = 1, \ldots, T_K = 1, k = 1, \ldots, K) \tag{7}
\]

which is the probability that an individual who is truly on layoff answers at least one the \( K \) compound questions negatively.

Next, we define the latent variable, \( W \), as the number of compound questions for which the true response is positive, i.e.,

\[
W = \begin{cases} 0 & \text{if } T_1 = 2, T_2 = 2, \ldots, T_k = 2 \\ 1 & \text{if } T_1 = 1, T_2 = 2, \ldots, T_K = 2 \\ \ldots \text{etc} \ldots & \text{if } T_1 = 1, T_2 = 1, \ldots, T_K = 1. \end{cases} \tag{8}
\]

For example, \( W = 0 \) if a person’s true response pattern to the questions O1–O4 is (2,2,2,2), \( W = 1 \) if the true response pattern is (1,2,2,2), and so on. Note that \( W = K \) corresponds to a true layoff. Thus, for the original questionnaire, \( W = 0, \ldots, 4 \) and for the revised questionnaire, \( W = 0, \ldots, 5 \).

To decomposing the probability in (7) into individual components for the compound question, \( Q_k \), we rewrite (7) in terms of the error probabilities associated with each compound question. Thus, it can be shown that (7) can be rewritten as

\[
\sum_{k=1}^{K} Pr(Q_1 = 1, \ldots, Q_{k-1} = 1, Q_k = 2 \mid W = K). \tag{9}
\]

The \( k^{th} \) term in the sum may be interpreted as the contribution of question \( Q_k \) to probability of being misclassified given a true LAYOFF.

To estimate the components of (9) using MLCA, we define a classification variable, \( R \), which is defined in analogy to \( W \) for the observed values of \( Q_k \), i.e.,

\[
R = \begin{cases} 0 & \text{if } Q_1 = 2, Q_2 = 2, \ldots, Q_K = 2 \\ 1 & \text{if } Q_1 = 1, Q_2 = 2, \ldots, Q_K = 2 \\ \ldots \text{etc} \ldots & \text{if } Q_1 = 1, Q_2 = 1, \ldots, Q_K = 1. \end{cases} \tag{10}
\]

\begin{table}
<table>
<thead>
<tr>
<th>Questionnaire</th>
<th>Question Wording</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original Questionnaire</strong></td>
<td></td>
</tr>
<tr>
<td>Q19</td>
<td>What were you doing most of LAST WEEK?</td>
</tr>
<tr>
<td>Q20</td>
<td>Did you do any work at all LAST WEEK not counting work around the house?</td>
</tr>
<tr>
<td>Q21</td>
<td>Did you have a job or business from which you were temporarily absent or on layoff LAST WEEK?</td>
</tr>
<tr>
<td>Q21A</td>
<td>Why were you absent from work LAST WEEK?</td>
</tr>
<tr>
<td>Q22E</td>
<td>Could you have taken a job LAST WEEK if one had been offered?</td>
</tr>
<tr>
<td><strong>Revised Questionnaire</strong></td>
<td></td>
</tr>
<tr>
<td>Q20</td>
<td>LAST WEEK, did you do ANY work (either) for pay (or profit)?</td>
</tr>
<tr>
<td>Q20B-a</td>
<td>LAST WEEK, (in addition to the business,) did you have a job, either full or part time? Include any job from which you were temporarily absent.</td>
</tr>
<tr>
<td>Q20B-b</td>
<td>LAST WEEK, were you on layoff from a job?</td>
</tr>
<tr>
<td>Q20B-1</td>
<td>What was the main reason you were absent from work LAST WEEK?</td>
</tr>
<tr>
<td>Q21</td>
<td>Has you employer given you a date to return to work?</td>
</tr>
<tr>
<td>Q21A-1</td>
<td>Have you been given any indication that you will be recalled to work within the next 6 months?</td>
</tr>
<tr>
<td>Q21A-2</td>
<td>Could you have returned to work LAST WEEK if you had been recalled?</td>
</tr>
</tbody>
</table>

**Figure 3.** Primary Components of UEM for the Original and Revised Questionnaires.
### Compound Questions Used in the LAYOFF Recode for Original and Revised Questionnaire Versions

<table>
<thead>
<tr>
<th>Compound Question Number</th>
<th>Source Question(s) from the CPS Questionnaire</th>
<th>Compound Question Response is Positive if Source Question Response is…</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>Q19: What were you doing most of LAST WEEK? or Q20: Did you do any work at all LAST WEEK not counting work around the house?</td>
<td>Q19: Any response except working and Q20: No</td>
</tr>
<tr>
<td>O2</td>
<td>Q21: Did you have a job or business from which you were temporarily absent or on layoff LAST WEEK?</td>
<td>Temporary layoff (Under 30 days) or Indefinite layoff (30 days or more or no definite recall date)</td>
</tr>
<tr>
<td>O3</td>
<td>Q21A: Why were you absent from work LAST WEEK?</td>
<td></td>
</tr>
<tr>
<td>O4</td>
<td>Q22E: Could you have taken a job LAST WEEK if one had been offered?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Let $\pi_{K|k}^{RW}$ denote $\Pr(R = k | W = K)$. Then for $k > 0$ we may write

$$\pi_{K|k}^{RW} = \Pr(Q_1 = 1, \ldots, Q_{k-1} = 1, Q_k = 2 | W = K).$$  \hspace{1cm} (11)$$

Thus, the contributions to error of each LAYOFF question can be obtained from the probabilities in (11).

To estimate the probabilities $\pi_{K|k}^{RW}$ we fit MLCA models to the same data from the 1993 and 1994 CPS as used in the previous analysis and replicated the analysis on the 1993 parallel survey data. Data from the 1992 and 1995 CPS were not part of this analysis. The MLCA models used were similar to those described in the analysis for Tables 1 and 2. That is, we used three consecutive months of data and estimated the components in (10) for 10 consecutive, overlapping intervals for each year (i.e., January–March, February–April, and so on to October–December). For the original questionnaire, the model specified three latent variables corresponding to the three months within a time period, each with $K + 1 = 5$ latent classes. For the revised questionnaire, we use an identical model except each latent variable had $K + 1 = 6$ latent classes.

As before, the best MLCA model for this analysis incorporated the proxy-self grouping variable, $P$, and specified non-stationary transitions, equal response probabilities within time period, group heterogeneous transition probabilities, and heterogeneous response probabilities. The model provides an adequate fit to the data for all months in the analysis (i.e., $p > 0.05$).

Table 3 provides a summary of the results from this analysis. In the column labeled “percent of total” we report $p_k \times 100$ percent where

$$p_k = \frac{\pi_{K|k}^{RW}}{\sum_{k=1}^{K} \pi_{K|k}^{RW}}.$$  \hspace{1cm} (12)$$

is the proportion of the classification error due to compound question $k$ in Figure 4 and where $\pi_{K|k}^{RW}$ are the MLCA estimates of $\pi_{K|k}^{RW}$.

The contribution to total error presented in Table 3 (Percent of Total column) is estimated by $p_k \times \Pr(A \neq 2 | X = 2)$ where $p_k$ is given by (12) and $\Pr(A \neq 2 | X = 2)$ is estimated from Table 2 as 1 minus the accuracy rate for LAYOFF. For the original questionnaire, the components that contribute most to LAYOFF classification error are question O2 (64.2 percent) and question O1 (27.2 percent). These two questions taken together explain more than 90 percent of the error in the LAYOFF classification.
For the revised questionnaire, estimates from the 1994 CPS indicate that more than 90 percent of the error in the LAYOFF classification arises from two components: N1 and N4.

The analysis for the revised questionnaire was repeated on the Parallel Survey with very similar results. The same two components emerge as contributing more than 90 percent of the error. As mentioned in section 2, the utility of the 1993 Parallel Survey as an indicator of data quality for the revised questionnaire is in doubt. Nevertheless, the agreement of the results from the Parallel Survey and the 1994 CPS adds strength to the findings from the 1994 CPS analysis.

Thus, reduction in LAYOFF classification accuracy for the revised questionnaire appears to be due primarily to error in the responses to two compound questions: N1, the revised global question “LAST WEEK, did you do ANY work (either) for pay (or profit)?” and N4, which determines whether an individual reporting some type of layoff has a date or indication of a date to return to work. The MLCA estimates indicate that almost 60 percent of the error in the revised LAYOFF classification maybe attributed to N1 while about 34 percent may be attributed to N4.

4.2.2 Decomposition of the LOOKING Recode

The estimation process described for LAYOFF was also applied to the LOOKING recode. Note that compound question O1, O2, N1, and N2 defined in Figure 5 for LOOKING are the same questions as defined in Figure 4 for LAYOFF. Since O1, O2, and N1 appeared to be problematic for LAYOFF, we might expect that they might also be problematic for LOOKING.

Following the approach used for LAYOFF, for each survey year, we defined a latent variable, \( W \) in (8) and an indicator variable, \( R \) in (9). As we did in the LAYOFF analysis, we fit MLCA models to the data and determined that the best MLCA model for the analysis is the model incorporating the proxy-self grouping variable, \( P \), and specifying non-stationary transitions, equal response probabilities within time period, group heterogeneous transition probabilities, and heterogeneous response probabilities. This model provides an adequate fit to the data for all months in the analysis (i.e., \( p > 0.05 \)). As before, we include the results from the Parallel Survey for comparison with the 1994 CPS results; however, the latter results will be emphasized.

Table 4 displays the values of \( p_k \) defined in (11) for the LOOKING classification. For the original questionnaire, the major contributors to classification error appear to be questions O1 and O3, which contribute 31.5 and 56.3 percent of total classification error, respectively. Question O2, which was quite problematic for the LAYOFF population, appears less so for the LOOKING population. While it contributes 64.2 percent of the LAYOFF error estimate (or 24.8 percentage points to the error rate), O2 only contributes 11.3 percent of the LOOKING error estimate (or 2.5 percentage points to the error rate).

For the revised questionnaire, the results from the analysis of the Parallel Survey and the 1994 CPS are again quite similar. The component N1 appears to be an important source of error for LOOKING as it was for the LAYOFF analysis. However, its contribution to LOOKING is smaller: 10 percentage points compared with 25 percentage points for LAYOFF. The biggest contributor to LOOKING error seems to be question N3 which contributes 64.5 percent of the error based on the CCO analysis and 51.1 percent based on the 1994 CPS analysis.

Thus, the initial labor force question appears to be problematic for both questionnaire versions. The MLCA suggests that persons who are looking for work as well as persons who are on layoff experience some difficulty responding to the question “LAST WEEK, did you do ANY work (either) for pay (or profit)?”. The changes made to this question in 1994 do not appear to have improved the accuracy of this question for the either population.

Table 3
Percent Contributions to Error in LAYOFF Classifications for Compound Questions for the 1993 CPS, Parallel Survey, and the 1994 CPS

<table>
<thead>
<tr>
<th>Question</th>
<th>1993 CPS (Original Version)</th>
<th>Parallel Survey (Revised Version)</th>
<th>1994 CPS (Revised Version)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Error Rate</td>
<td>Percent of Total</td>
<td>Error Rate</td>
</tr>
<tr>
<td>O1</td>
<td>10.53</td>
<td>27.20</td>
<td>–</td>
</tr>
<tr>
<td>O2</td>
<td>24.84</td>
<td>64.19</td>
<td>–</td>
</tr>
<tr>
<td>O3</td>
<td>2.35</td>
<td>6.08</td>
<td>–</td>
</tr>
<tr>
<td>O4</td>
<td>0.67</td>
<td>1.74</td>
<td>–</td>
</tr>
<tr>
<td>N1</td>
<td></td>
<td></td>
<td>23.19</td>
</tr>
<tr>
<td>N2</td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>N3</td>
<td></td>
<td></td>
<td>2.76</td>
</tr>
<tr>
<td>N4</td>
<td></td>
<td></td>
<td>18.42</td>
</tr>
<tr>
<td>N5</td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>Total</td>
<td>38.39</td>
<td>100.00</td>
<td>44.37</td>
</tr>
</tbody>
</table>
Table 4
Percent Contributions to Error in LOOKING Classifications by Compound Questions for the 1993 CPS, Parallel Survey, and the 1994 CPS

<table>
<thead>
<tr>
<th>Question</th>
<th>1993 CPS (Original Version)</th>
<th>Parallel Survey (Revised Version)</th>
<th>1994 CPS (Revised Version)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Error Rate</td>
<td>Percent of Total</td>
<td>Error Rate</td>
</tr>
<tr>
<td>Old Questionnaire</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O1</td>
<td>6.93</td>
<td>31.51</td>
<td>–</td>
</tr>
<tr>
<td>O2</td>
<td>2.49</td>
<td>11.34</td>
<td>–</td>
</tr>
<tr>
<td>O3</td>
<td>12.39</td>
<td>56.33</td>
<td>–</td>
</tr>
<tr>
<td>O4</td>
<td>0.18</td>
<td>0.83</td>
<td>–</td>
</tr>
<tr>
<td>New Questionnaire</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N1</td>
<td>–</td>
<td>–</td>
<td>8.38</td>
</tr>
<tr>
<td>N2</td>
<td>–</td>
<td>–</td>
<td>0.00</td>
</tr>
<tr>
<td>N3</td>
<td>–</td>
<td>–</td>
<td>16.38</td>
</tr>
<tr>
<td>N4</td>
<td>–</td>
<td>–</td>
<td>0.46</td>
</tr>
<tr>
<td>N5</td>
<td>0.18</td>
<td>0.71</td>
<td>0.14</td>
</tr>
<tr>
<td>Total</td>
<td>22.00</td>
<td>100.00</td>
<td>25.39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Compound Question Number</th>
<th>Source Question(s) from the CPS Questionnaire</th>
<th>Compound Question Response is Positive if Source Question Response is….</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old Questionnaire</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O1</td>
<td>Q19: What were you doing most of LAST WEEK?</td>
<td>Q19: Any response except working and Q20: No</td>
</tr>
<tr>
<td></td>
<td>or Q20: Did you do any work at all LAST WEEK</td>
<td></td>
</tr>
<tr>
<td></td>
<td>not counting work around the house?</td>
<td></td>
</tr>
<tr>
<td>O2</td>
<td>Q21: Did you have a job or business from which you were temporarily</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>absent or on layoff LAST WEEK?</td>
<td></td>
</tr>
<tr>
<td>O3</td>
<td>Q22: Has … been looking for work during the past 4 weeks?</td>
<td>Q22: Yes or response to Q19 was LK (LOOKING) and Q22A: Response other than “nothing”</td>
</tr>
<tr>
<td></td>
<td>and Q22A: What has … been doing in the last 4 weeks to find work?</td>
<td></td>
</tr>
<tr>
<td>O4</td>
<td>Q22E: Could … have taken a job LAST WEEK if one had been offered?</td>
<td>Yes or No, and reason is “Already has job” or “Own temporary illness”</td>
</tr>
<tr>
<td>New Questionnaire</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N1</td>
<td>Q20: LAST WEEK, did you do ANY work (either)</td>
<td>Q20: No</td>
</tr>
<tr>
<td></td>
<td>for pay (or profit)?</td>
<td></td>
</tr>
<tr>
<td>N2</td>
<td>Q20B-a: LAST WEEK, (in addition to the business,) did you have a job,</td>
<td>Q20B-a:No¹</td>
</tr>
<tr>
<td></td>
<td>either full or part time? Include any job from which you were temporarily</td>
<td></td>
</tr>
<tr>
<td></td>
<td>absent.</td>
<td></td>
</tr>
<tr>
<td>N3</td>
<td>Q22: Have you been doing anything to find work during the last 4 weeks?</td>
<td>Yes</td>
</tr>
<tr>
<td>N4</td>
<td>Q22A: What are all the things you have done to find work during the last 4 weeks?</td>
<td>Mention of at least 1 active activity.</td>
</tr>
<tr>
<td></td>
<td>Or Q22A-DK: You said you have been trying to find work. How did you go about looking? And Q22A-DK1: Can you tell me more about what you did to search for work?</td>
<td></td>
</tr>
<tr>
<td>N5</td>
<td>LAST WEEK, could you have started a job if one had been offered?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

¹ Note: In a few cases, N2 was positive if response to Q20B-a was “Disabled” or “Unable” and response to Q20A-1: “Does your disability prevent you from accepting any kind of work during the next six months?” was “No”.

Figure 5. Compound Questions Used in the LOOKING Recode for Original and Revised Questionnaire Versions.

The key difficulty for the LOOKING category appears to be determining whether persons who are truly looking for work have made efforts of any type (either passive or active) in the past four weeks to find work. If a respondent is classified correctly as having made some effort, the next step in the process – viz., determining whether the efforts satisfy the definition of active looking – is not problematic according to the estimates in Table 4.

5. Conclusions

Biemer and Bushery (2000) provides some evidence that unemployment classification accuracy rates in the 1994 CPS redesign survey were smaller than for the original survey design used prior to 1994. This paper provides additional evidence of their findings based upon a more extensive analysis of CPS data from 1992 through 1994. Our results
indicate that the probability of correctly classifying unemployed persons decreased from 79.1 percent to 73.5 percent – a difference of 5.6 percentage points. We estimate that roughly 60 percent of the reduction (3.4 percentage points) is due to an increase in the classification error for persons on layoff while the remainder (2.2 percentage points) is due to an increase in the classification error for persons looking for work.

For the revised questionnaire, both LAYOFF and the LOOKING classifications are each based upon five compound questions. For LAYOFF, two compound questions emerged as being problematic. One is the initial labor force question, which asks “LAST WEEK, did you do ANY work (either) for pay (or profit)?” The contribution of this component to LAYOFF misclassification is estimated to be approximately 57 percent which is more than double the corresponding rate for this question in the original questionnaire. In addition, a large error rate is estimated for the compound question formed by two questions: “Has your employer given you a date to return to work?” and “Have you been given any indication that you will be recalled to work within the next 6 months?” Approximately 34 percent of the estimated LAYOFF error rate is due to this combination. Since there are no corresponding questions in the original questionnaire, most of the error in classifying persons on layoff in the revised questionnaire may be linked to these two questions.

For classifying persons who are looking for work in the redesign survey, two questionnaire components appear to contribute most to classification error: “LAST WEEK, did you do ANY work (either) for pay (or profit)?” and “Have you been doing anything to find work during the last 4 weeks?/What has...been doing in the last 4 weeks?” The error rates for both questions are slightly larger for the revised questionnaire than for the original questionnaire. These increases, therefore, explain the slight increase in LOOKING classification error observed for the revised questionnaire.

The error in CPS unemployment classification is well-documented; for example, see Chua and Fuller 1987; Abowd and Zellner 1985; Porterba and Summers 1995; and Sinclair and Gastwirth 1998. A widely accepted measure of reliability for the CPS – viz., index of inconsistency computed CPS reinterview – shows the reliability of the CPS unemployment classification decreased after the redesign. Results provided in this paper are consistent with these prior studies and help determine the source of the error in the CPS classification of the unemployed. At a minimum, our results provide a basis for further investigation into the root causes of the errors in the collection of labor force data in the CPS. Through cognitive laboratory experiments and field experiments, we may identify causes of the error in the unemployment questions that would suggest ways to improve the questions. Such improvements could be implemented in a future redesign of the CPS.

Acknowledgement

The author would like to acknowledge the assistance of Pamela McGovern at the U.S. Census Bureau who commented on early drafts of the paper. Appreciation is also expressed to the Associate Editor and an anonymous referee, both of whom were very helpful in preparing the article. Financial support for this research was provided by the U.S. Census Bureau.

References


Biemer, P., and Bushery, J. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data. Survey Methodology, 26, 139-152.


1. Introduction

I enjoyed very much reading this very well written paper. The topic addressed by Paul Biemer – classification errors in the measurement of employment status – is a very important one. Employment statistics belong to the most important macro-economic indicators and, actually, we would wish they would be free of error. It, however, turns out to be impossible to measure a person’s employment without error. The best that can be done is design the data collection in such a manner that the classification errors at the individual level are minimized as much as possible. The current paper contributes to this objective.

An earlier study by Biemer and Bushery (2000) indicated that the 1993 changes in the measurement procedure that intended to reduce classification errors actually increased measurement error. In the current paper, Paul Biemer replicates these former analyses with a longer time series and with an extra employment category obtained by splitting the unemployed group into “on layoff” and “looking for work”. The reported results confirm the earlier conclusions that the new procedure is worse than the old procedure. In a second step, Biemer tries to disentangle the sources of measurement error for the two unemployed categories by modeling the separate questions that are used to determine whether a person is “on layoff” and “looking for work”, respectively. Sources of error are identified that point at possible improvements in the questionnaire.

Because of my background, my commentary will mainly concern methodological and statistical issues. More precisely, I will discuss some methodological problem related to application of the LC Markov model, as well as indicate how the statistical analysis could be somewhat refined. I am much more concerned about the third assumption; that is, the assumption of independent classification errors (ICE) over time (Bassi, Hagenaars, Croon and Vermunt 2000). Is it realistic to assume that the occurrence of a certain type of classification error at time point \( t \) does not affect the probability of making the same mistake at time point \( t + 1 \)? In my opinion, this assumption is not realistic in the current application. For example, a respondent who makes a mistake because (s)he did not understand one of the questions will most probably (or at least be more likely than others) make the same error again at the next occasion. In my opinion, it is necessary to conduct a simulation study to determine the sensitivity of the estimated classification errors for violations of the ICE assumption.

I have another critical remark concerning the use of the LC Markov model for quantifying measurement error in a person’s employment state. According to the model, there is a probabilistic relationship between an individual’s true and observed states. What is, however, the true state? Is it the true employment state occupied at a particular time point, or the state that would have been recorded with an error-free or gold-standard instrument? Or is it the state a person would have occupied under “normal conditions”? That is, if also randomness in his/her behavior is filtered out.

I will illustrate my point with a small example. Suppose that there is two types (two latent segments) of coffee consumers: consumers who prefer brand A and consumers who prefer brand B, and that I belong to the brand B segment, which means that under normal circumstances I buy brand B coffee. In an interview, I am asked which

---

brand I bought last week. Suppose I report that I bought a brand A package of coffee, and that am neither lying nor making a mistake. In other words, there is no classification error in the sense of making a mistake: I really bought brand A this week (the researcher doesn’t know that of course). On the other hand, my behavior from this week is inconsistent with my preference, which means that in terms of measurement of my preference there is a classification error. This example illustrates that there are two types of “errors” that can be made: an error in the reporting and an “error” in the behavior. The “error” in my behavior of this week may have many causes, such as “brand B was sold out”, “brand A was offered at a lower price this week”, “I could not find the brand B package because of changes in the arrangement of the supermarket”, etc. The LC Markov model is not able to distinguish such randomness in the behavior that is uncorrelated across time points from real classification errors.

What does this imply for the employment application? It implies that an individual’s true state may be “on layoff”, but for some reason (by chance) this particular month (s)he has worked. If this “some reason” is uncorrelated with other “some reasons” for being in the “wrong” observed state at other occasions, it will be labeled classification error by the LC Markov model. While in the case of the measurement of preferences based on revealed (or stated) preferences correcting for randomness in behavior seems to be exactly what we wish to accomplish, this is clearly not the case in the measurement of employment status. I, therefore, have the strong feeling that the error rates reported by Biemer might be somewhat overestimated because of randomness in employment behavior, for instance, caused by randomness in the functioning of the labor market.

A well-known consequence of modeling individual change by means of a LC Markov model is that the estimated number of latent transitions is much smaller that the corresponding observed numbers. The reason for this is that both independent classification errors and independent random behavior is filtered out; that is, part of the observed change is attributed to these phenomena.

I would have set up the model in a somewhat more elegant and less ad hoc manner. Instead of running a separate analysis for each of the rotation groups, I would have tried to build a simultaneous model for all rotation groups. The main problem of doing a series of separate analyses is that parameters that should actually be equated across rotation groups are now estimated without constraints. For example, the employment distribution in March 1994 should be the same in the rotation groups that were interviewed between January and March, February and April, and March and May, respectively. Moreover, the transition probabilities between March and April should be the same in the February–April and March–May rotation groups. This has also implications for the Parallel Survey groups: their time-specific latent distributions and transitions should be assumed to be equal to the ones of the standard CPS. That would have been a much better manner to test whether measurement error differ between the two questionnaires. Especially for the period in which the questionnaire forms overlap, it is crucial to assume equal latent distributions in order to be able to prevent that differences in measurement error appear partially as differences in true states.

A similar problem of the separate analyses applies for the estimation of the classification errors. These are assumed to be time-constant within the 3–month period that a rotation group is interviewed, but are allowed to differ across rotation groups, even if they are interviewed in the same month. It would, of course, be much better to impose equality constraints across rotation groups. A consistent application of the time-homogeneity assumption would imply that – both for the old and the new questionnaire form – the measurement errors are constant within the full investigation period.

What we, actually, need is a LC Markov model covering all 30 months; that is, a model for 30 instead of 3 time points. Such a simultaneous model for all rotation groups is as easily specified as a model for 3 time points. Of course, for each rotation group, only 3 of the 30 months are observed, which means that the other time points have to be treated as missing values. This is not a problem in the maximum likelihood estimation of the model parameters since we can simply assume that the data are missing at random (Vermunt 1997). Questionnaire type (old/new) serves as grouping variable (in addition to interview mode) and affects the time-homogenous classification error probabilities. In other words, we estimate only two sets of classifications errors, one for the old and one for the new questionnaire. Transition probabilities may change over time, but will be equal across rotation groups interviewed at the same occasions. Moreover, the initial state probabilities of a rotation group are not estimated as separate parameters.
since they are defined by the current state of the latent Markov chain.

A practical problem of the simultaneous modeling is that with so many time points it is no longer possible to estimate the model parameters with the standard EM algorithm. With a variant of EM called the Baum-Welch algorithm, however, the model can also be applied with many time points (Vermunt 2003; Paas, Bijmolt and Vermunt 2003). This algorithm is implemented in an experimental version of the Latent GOLD program (Vermunt and Magidson 2000, 2003) and will be available in a next version of this program.

An alternative way to implement a simultaneous model is as a LC Markov model for 3 occasions in which rotation group serves as grouping variable and in which the relevant across rotation group equality restrictions are imposed on the classification errors, transition probabilities, and initial state probabilities. The most complicated part of this approach is that it requires the use of restrictions on marginal probabilities (Vermunt, Rodrigo and Ato-García 2001). More precisely, the initial state probabilities should be in agreement with the marginal class sizes in the rotation groups that are interviewed at the same occasion.

Other aspects of the modeling that could be refined are the treatment of missing values and the coding of the interview mode. It is not necessary to eliminate cases with missing values from the analysis as is done by Paul Biemer because ML estimation with missing values is straightforward. As far as the interview mode is concerned, it would be much more elegant to work with only two categories – proxy and self – instead of four categories and let the interview mode vary across occasions within cases. In other words, interview mode could be used as a time-varying covariate. Vermunt, Langeheine and Böckenholt (1999) proposed such a latent class Markov model with time-varying covariates.

4. Model for Response Process

It is a very nice idea to try to disentangle which questions in the questionnaire are causing the classification errors by modeling the response process itself. This may yield lots of valuable information for redesigning the questionnaire. I, however, think that the extended models for the employment statuses “on layoff” and “looking for work” are formulated in an overly complicated manner.

The form of the created variable \( R \) is the same as of the outcome variable in a sequential choice analysis or in a discrete-time survival analysis. Answering the next question is fully determined by whether the current one is answered positively or not. The information we have is how many steps a person takes, which is conceptually equivalent to a discrete survival time. A person “surviving” till the end is classified as being “on layoff” (“looking for work”).

In my opinion, it is not very helpful to treat this variable as being generated by \( K \) latent variables (Ts). This only makes sense if theoretically there should be a response hierarchy at the latent level, which, however, because of measurement error, is not encountered at the manifest level. That is, if at the manifest level there are \( 2^K \) instead of \( K \) possible responses. Even if it is the case, it often suffices to conceptualize the model as a model with a latent variable with \( K + 1 \) classes and \( K \) indicators, a structure that is sometimes referred to as a probabilistic Guttman model.

Paul Biemer recognizes the complexity of the \( K \) latent and \( K \) manifest variables formulation and decides to simplify the model. However, I assume because of his starting point, he decided to keep \( K + 1 \) latent classes. I do not see why so many latent classes are needed. There are not even so many employment states. More logical would be to have only two classes – “on layoff” and “not on layoff” (“looking for work” and “not looking for work”) – since the questions are only intended to make this particular distinction. It can, of course, happen that the questions turn out to be informative about the type of “not on layoff” (“not looking for work”) status, in which case an extra latent class might be needed. What is clear to me is that \( K + 1 \) classes are far too many.

I was wondering how many persons were classified as “on layoff” (“looking for work”) at the various time points in the analysis with composite variable \( R \) as indicator. Are these numbers, as well as the number of transitions into and out off this state similar to the ones obtained with the standard four-state LC Markov model. In my opinion, this is a requisite for the validity of the calculation performed to obtain the figures presented in Tables 3 and 4.

A final thing that occurred to me is the following. Why not building a LC Markov model using the full questionnaire information as is done in the second part of the analysis. In other words, an alternative to using the observed constructed classification consisting of 4 employment categories would be to use the full set of CPS employment questions answered by the respondents. Such an analysis with multiple indicators would not only be much more informative, it would also make it possible to test and relax some of the assumptions that were made in the current analysis. For example, the ICE assumption could be relaxed for some of the questionnaire items.

References


1. Introduction

We are grateful for the opportunity to comment on this interesting paper. We will focus most of our comments on the empirical findings about the 1994 Current Population Survey (CPS) redesign, rather than a technical discussion of the Markov Latent Class Analysis (MLCA) methodology itself.

In his article, “An Analysis of Classification Error for the Revised Current Population Survey Employment Questions,” the author applies MLCA models in an effort to trace the source of what he believes to be the “reduced accuracy of the revised classification of unemployed persons” after the redesign. In the CPS individuals are considered to be unemployed either because they are classified as being on layoff or because they are classified as looking for work. The author reports a particularly large reduction in the accuracy of the measurement of persons on layoff. Consequently, we will focus our attention on the classification of individuals on layoff, although similar comments can be made about the change in the measurement of those looking for work. In examining the accuracy of the measurement of those on layoff, the author assumes that those classified as on layoff were conceptually the same before and after the 1994 redesign, and that these individuals should exhibit the same labor force flows month-to-month. There are, however, many reasons why the improved measurement embodied in the redesign should conceptually change who is classified as on layoff. In addition, there are several factors unrelated to changes in question wording that could affect the composition of those classified as on layoff. Therefore, what the author describes as a reduction in accuracy due to the redesign more appropriately could be attributed to conceptual changes in those classified as on layoff, and the fact that what was being measured by the CPS before the redesign is not the same as what is being measured by the CPS after the redesign.

2. Improved Measurement

One of the main reasons for the CPS redesign was to more accurately measure official definitions and concepts. Layoff was found to be an especially problematic concept, in that its meaning in general usage in the 1990’s – a permanent job separation – was very different from the official CPS definition – a temporary job separation with the expectation of recall. When the questions were originally written in the 1940’s, the term layoff was commonly used to refer to temporary spells of unemployment due to retooling or slowing of business conditions. Consequently, recall expectations were not asked about in the pre-redesign questionnaire. Research conducted in the 1980s and early 1990s in preparation for the redesign indicated that respondents’ interpretation of layoff had become considerably broader than the official definition. Focus group interviews and large scale respondent debriefings found that between 30 and 50 percent of those who said they were on layoff did not expect to return to their former employers (Rothgeb 1982; Palmisano 1989; Polivka and Rothgeb 1993) Also, in 1993, 5.4 percent of those classified as on layoff had last worked 1 to 5 years ago, and another 0.6 percent had not worked in the last 5 years. This lack of recent work experience further supports the notion that many of those classified as on layoff prior to the redesign had no expectation of recall.

To better measure the official CPS definition of layoff, two questions were added in the revised questionnaire asking about individuals’ recall expectations – “Has your employer given you a date to return to work?” and “Have you been given any indication that you will be recalled to work within the next 6 months?” Individuals for whom the answer is “yes” to either of these questions are classified as on layoff if they are available for work; all others are excluded from being classified as on layoff (these individuals can be classified as unemployed later in the questionnaire if they meet the active job search and availability criteria).

As a result of the addition of these direct questions, a somewhat different group of people would be expected to be classified as on layoff. Prior to the redesign, a substantial proportion, if not the majority, of individuals classified as on layoff were in fact permanently separated from their employers. After the redesign, those classified as on layoff had to expect to be recalled to their former employers; thus the vast majority of these individuals should be only temporarily separated from their employers. It is not
surprising that these two groups of individuals would exhibit different month-to-month flows between labor force groups. It is reasonable to expect that individuals who expect to be recalled to their job would be more likely than those who are permanently separated to go from being temporarily on layoff to employed in consecutive months. Further, compared with permanently separated workers, those in industries in which temporary layoffs are prevalent would be more likely to be on layoff one month, employed the next month, and then laid off again.

Month-to-month gross flows of individuals between labor force states indicate that there was an increase in the proportion of the unemployed who went to employment after the 1994 redesign. Specifically, in 1994, 26.6 percent of those who were unemployed in the first month were employed in the second month, compared with 23.7 percent in 1993.

The author’s MLCA estimates of a supposed decrease in the accuracy of those classified as on layoff after the redesign because more individuals are classified as employed subsequent to being on layoff, in reality is exactly in accord with what would be expected with a tightening of the definition of on layoff, and is consistent with the increase in the month-to-month gross flows between unemployment and employment (although the increased flow also is in accord with a declining unemployment rate that was observed during the time period covered by the author’s study). The MLCA’s smaller, but still significant, estimated decrease in accuracy due to more individuals on layoff being classified as not in the labor force after the redesign also is consistent with the tightening of the definition of on layoff through the requirement that individuals expect to be recalled in the next six months, given that individuals may adapt or change their recall expectations over time. For instance, when first interviewed, individuals may expect to be recalled in the next six months. However, in subsequent months, as the time from the initial separation increases, these individuals may no longer say that they expect to be recalled. If, at the same time, these individuals have not started searching for alternative employment, perhaps because they are still eligible to receive unemployment insurance payments, these individuals would transition to being not in the labor force. Alternatively, individuals may initially expect to be recalled; however, in subsequent months due either to poor weather conditions or a deteriorating economic situation for their former employers these individuals may become more uncertain about the probability of being recalled and thus they may not say that they expect to be recalled. If in later months, economic conditions for their former employers improve or the weather becomes less inclement, these individuals again may correctly feel that they will be recalled. The existence of changing expectations could generate a three month pattern where individuals truly were on layoff in the first month, not in the labor force the second month, and on layoff again in the third month. Those who were permanently separated from a job and were incorrectly classified as on layoff in the redesigned survey would be unaffected by changing recall expectations. Consequently, individuals who were permanently separated from their jobs probably would be more likely to report themselves as on layoff in consecutive months with the redesigned survey. The MLCA model would interpret this greater stability as indicating that those on layoff were more accurately measured prior to the redesign. However this greater “accuracy” would only be amongst those who were incorrectly classified because they used too broad a definition.

The author concludes that 60 percent of the misclassification of those on layoff in the redesigned survey is due to the question “LAST WEEK, did you do ANY work for pay?” This actually is consistent with more people being on temporary layoff and being recalled by their former employers in the redesigned survey (although if individuals on layoff engage in temporary employment while waiting to be recalled to their former employers, an increase in transitions to employment after 1994 may also be at least partially attributable to the broader employment question used in the redesigned survey). Similarly, the author concludes that 40 percent of the misclassification of those on layoff in the redesigned survey is due to the expectation of recall questions (“Has your employer given you a date to return to work?” and “Have you been given any indication that you will be recalled to work within the next 6 months?”). This is consistent with changing recall expectations and a slight increase in the flow between on layoff and not in the labor force. The author is obtaining different MLCA estimates of those classified as on layoff before and after the redesign because the composition of those groups has been changed, and the composition of the groups have changed in a manner that was desired and intended by those who redesigned the questionnaire.

Further evidence of the different composition of those classified as on layoff can be found in a comparison of data that were collected to determine the effect of the redesign on labor force estimates generated from the CPS. Prior to January 1994, the redesigned questionnaire was administered to 12,000 households monthly from late 1992 to December 1993. After the new questionnaire was implemented in 1994, the old questionnaire was administered monthly from January 1994 to May 1994 to 12,000 households drawn from the same sample. The experimental administration of the old and redesigned questionnaires has been referred to as the “Parallel Survey”. Parallel Survey estimates from before 1994 using the new methodology and
after 1994 using the old methodology were generated to compare to official CPS estimates using the unrevised CPS procedures prior to 1994 and the redesigned procedures after 1994. Polivka and Miller (1998) illustrate the importance of using both parts of the Parallel Survey to obtain a complete picture of the effects of the redesigned survey. For instance, if just the first part of the Parallel Survey was used, it would have been estimated that the redesign increased the unemployment rate by 0.5 percentage point. In fact, when both parts of the Parallel Survey were used, the redesign was estimated to have no statistically significant effect on the unemployment rate.

Using both parts of the Parallel Survey and the official CPS estimates, Polivka and Miller estimate that the redesigned CPS decreased the proportion of unemployed men who were on layoff by a little less than 7 percent, while it increased the proportion of unemployed women classified as on layoff by almost 7 percent (although the latter estimate was not statistically significant at a 5 percent level). These estimates imply that the redesign would decrease the proportion of those on layoff who were male and increase the proportion who were female compared to the proportions that were obtained prior to the redesign, if all else were equal. Comparison of annual averages for those over the age of 20 support this notion, since they indicate that, in 1993, 67.2 percent of those on layoff were male, compared to 63.6 percent of those on layoff in 1994 (although in addition to questionnaire changes these proportions could be affected by changes in economic conditions).

The industry distribution of those classified as on layoff, using data from both parts of the Parallel Survey and the official CPS, reveals other compositional changes in those classified as on layoff before and after the redesign. Examination of estimates from the redesigned survey to the official CPS estimates for January to May 1993 and from the unrevised survey to official CPS estimates for January to May 1994 reveals particularly dramatic differences for those in the durable manufacturing industry. The proportion of those on layoff who were formerly employed in durable manufacturing when the unrevised questions were used was almost half the proportion obtained when the redesigned questions were used (for January to May 1993 the proportion of those on layoff who were formerly employed in durable manufacturing averaged 16.8 percent among those who received the unrevised questions and 9.8 percent among those who received the redesigned questions. For January to May 1994 the proportions were 8.7 percent among those who received the unrevised questions and 15.5 percent for those who received the redesigned questions). At the same time the proportion of those on layoff who were in construction was 10 to 15 percent larger when the redesigned questions were used compared to when the unrevised questions were used (for January to May 1993 the proportion of those on layoff who were formerly employed in the construction industry averaged 33.3 percent for those who received the redesigned questions and 27.4 percent for those who received the unrevised questions. For January to May 1994 the proportions were 33.3 percent and 25.9 percent respectively).

Averaging the average difference between the first part of the Parallel Survey and the CPS for January 1993 to May 1993 (which is equal to the new method effect minus the Parallel Survey effect) with the average difference between the CPS and the second part of the Parallel Survey for January 1994 to May 1994 (which is equal to the new method effect minus the Parallel Survey effect) indicates that the redesign decreased the proportion of those classified as on layoff who were formerly employed in the durable manufacturing industry by 7.3 percentage points and increased the proportion classified as formerly employed in the construction industry by 3.7 percentage points (averaging the average difference between the first part of the Parallel Survey and the CPS with the average difference between the CPS and the second part of the parallel survey is in the spirit, albeit a simplified version, of the main-effects linear models estimates using generalized least squares that were presented in Polivka and Miller).

Individuals in different industries could have very different true labor force transition patterns which in turn could be influencing the MLCA estimates. For instance, given that a substantial proportion of employment in the construction industry is sensitive to weather conditions and may be more project-oriented than other types of employment, it is not unreasonable to expect that workers in construction might truly be more likely to be temporarily laid off in the first of three consecutive months, employed on a short term basis in the second month (either because the weather improved in the second month or because a short term construction project was undertaken), and then temporarily laid off again in the third month (either because weather conditions deteriorated or the project for which they were hired was completed). On the other hand, employment in the durable manufacturing industry has been steadily declining since the 1970’s (for example, comparing non-recession years, it was estimated that in 1971 14.9 percent of U.S. workers as measured by BLS’s establishment survey were employed in the durable manufacturing industry, compared to 9.2 percent in 1993 and 8.5 percent in 2000). This long term decline in employment makes it likely that a large proportion of workers in the manufacturing industry classified as “on layoff” prior to the redesign were permanently separated from their employers (the change in the industry distribution when the expectation of being recalled was imposed is consistent with this notion). Being
permanently separated from a job in combination with the relatively high wages workers in durable manufacturing received may increase the likelihood of these individuals being unemployed in three consecutive months, because it takes time to find employment in another industry at a similar wage.

Comparison of MLCA model estimates before and after the redesign without accounting for differences in industry composition of those classified as on layoff could cause analysts to mistakenly conclude that the redesign decreased the accuracy of labor force classifications. In reality, the increase in transitions that were measured after the redesign represented a true increase in transitions to employment after layoff was properly asked about in the CPS questionnaire. Failure to account for the fact that the redesigned CPS questionnaire intentionally classified a somewhat different group of individuals on layoff than did the unredesigned questionnaire could lead to incorrect conclusions being drawn from the MLCA models. Workers permanently separated from their employers who were classified as on layoff using the unredesigned questions are appearing to be more accurately classified in MLCA models, but they are more stable in a classification that was incorrect in the first place. Further, a proportion of individuals who are correctly classified as on layoff according to the official definition inherently could have less stable employment histories due either to their personal tastes or the industries with which they are associated.

In addition to compositional changes related to differences in question wording, the author also may have inadvertently captured in his estimates several other compositional changes unrelated to wording differences. These include differences in the time periods the author used for his estimates, as well as technological changes in the data collection process and economic conditions.

3. Seasonality

The first inadvertent compositional difference the author may have introduced is related to seasonality and the different time frames the author used for estimation. The number of individuals classified as on layoff in the CPS has a great deal of seasonal variability, with typically a larger number of individuals being on layoff early in the year. For instance, there were 358 individuals who were classified as on layoff in January 1995 who matched to February and March, while there were 294 individuals classified as on layoff in March 1995 who matched to April and May, and only 188 people classified as on layoff in June 1995 who matched to July and August. This means that there were 18 percent more people initially classified as on layoff in January 1995 than in March 1995 and 47 percent more individuals classified as initially on layoff in January 1995 than in June 1995. Using three month moving averages generated with the same calendar months probably would help to mitigate the effects of seasonality. However, the author did not use the same monthly time spans to generate his three-month moving averages to estimate the MLCA models before and after the redesign. The majority of the author’s pre-redesign estimates were generated using data from August 1992 through December 1993, while the majority of his post-redesign estimates were generated using data from January 1994 to May 1995. Using these time spans means that the author only has, for instance, one January to March matched set of data for the pre-redesign estimates, while he has two January to March matched sets of data for the post-redesign estimates.

4. Technological Changes in Data Collection

A second reason that the composition of the groups in various labor force states may be different for data collected with the unredesigned and the redesigned methodology is related to the ability to match individuals’ data from month to month and the quality of these matches. The vast majority of data collected using the unredesigned methodology either in the official CPS prior to January 1993 or in the Parallel Survey from January 1994 to May 1994 were recorded using a paper form, and interviewers were required to transcribe by hand household and person identification numbers from master files to the paper survey forms. All of the data collected using the redesign methodology, either in the official CPS after January 1994 or in the Parallel Survey in 1993, were collected using an automated instrument that was loaded onto either a laptop computer or on a centralized computer. As part of the computerized data collection process, household and person identification numbers were automatically and consistently carried forward month to month. Using paper forms and transcribing data by hand has the potential to introduce errors and cause researchers to eliminate as non-matches individuals who actually are the same individuals and thus true matches.

Using the same public-use data that the author used, in combination with additional information about whether an individual had moved (that is periodically collected in the CPS), Madrian and Lefgren (1999) estimated that, depending on the stringency of the match criterion used, between 64 and 87 percent of those who were eliminated as an invalid match probably legitimately did match. Further, Madrian and Lefgren noted that there was a substantial decline between 1993 and 1996 in the fraction of invalid matches that probably should have been retained in the data set based on the criterion of whether an individual had
moved (since Madrian and Lefgren were using publicly released data, they were not able to investigate the validity of matches for 1994 to 1995 and 1995 to 1996 because the ability to match this data was suppressed to protect individuals’ confidentiality). Madrian and Lefgren suggest that the increased number of valid matches for 1996 onward was due to improvements attributable to the redesign (it should be noted that, although a better match can be obtained using data internal to BLS and the Census Bureau in which information has not been suppressed, the quality of a match using internal data still will be affected by the data collection methodology. Thus the quality of the match will be better after the redesign than before the redesign). In their research, Madrian and Lefgren also found that individuals who were incorrectly excluded from the matched data sets were much more likely to be young and have their information provided by another member of the household (a proxy responder). These individuals are also the ones that Biemer argues are more likely to have classification errors in their labor force status. Consequently, by potentially including more of these individuals in his study due to the improved quality of the match, the author could be obtaining a decrease in the accuracy of his measures that he incorrectly is attributing to the questionnaire.

5. Economic Conditions

Economic conditions may also contribute to differences in the composition of the groups classified as on layoff before and after the redesign. From 1992 to 1995, the period which the author uses for the majority of his MLCA modeling, the unemployment rate was steadily declining. Specifically, in 1992 the annual average unemployment rate was 7.5 percent while in 1995 it was 5.6 percent.

At a higher unemployment rate, it is likely that the proportion of individuals who remain unemployed month to month is larger than at lower unemployment rates. As the economy improves and the unemployment rate declines, it is not unreasonable to expect an increase in the proportion of individuals who transition from being on layoff to employment. With the increase in these transitions to employment, the proportion of individuals who transition to temporary jobs might also increase. Indeed, although undoubtedly related to many factors, the number of individuals employed in the temporary help supply industry (as defined under the NAICS coding system) increased 44 percent between 1992 and 1995 – from 1.1 percent to 1.5 percent of the U.S. establishments’ payrolls (as measured by the BLS’s establishment survey).

In addition, as the unemployment rate declines, the type of individual classified as unemployed may change. Specifically, those who remain unemployed when the unemployment rate is low tend to find it more difficult to become steadily employed and are more likely to transition quickly between labor force states. This is the logic behind studies that analyze the effects of different types of employment separations on subsequent labor force outcomes. For instance, in a study comparing individuals who were separated from their employers due to slack business conditions as opposed to complete plant shut downs, Gibbons and Katz (1991) found that, with regard to both duration of joblessness and earnings, workers who were separated from their employers due to slack business conditions did significantly worse than did those who were separated due to a plant closing. Gibbons and Katz argue that these differences were due to employers being able to dismiss their least productive workers, while retaining their more productive workers, when business conditions were slack, as opposed to employers having to dismiss both their least productive and most productive workers when a plant was completely shut down. Similarly, Darby, Haltiwanger and Plant (1985) argue that as economic conditions worsen, the duration of unemployment increases as a result of a change in the composition of those who are unemployed. This is because in more adverse economic conditions, the proportion of the unemployed who are high-skill workers (who also are less used to being unemployed and more likely to be able and willing to hold out for a more satisfactory job) will increase and the proportion of the unemployed who are less skilled and who frequently transition between labor force states will decrease.

It is important to note that the majority of the author’s pre-redesign estimates were generated using 1992 and 1993 data, when the unemployment rate averaged 7.0 percent, while the majority of the redesigned estimates were generated using data from 1994 and 1995, when the unemployment rate averaged 6.0 percent. Changes in general economic conditions, and corresponding changes in the composition of the unemployed, may be affecting the supposed accuracy of the author’s estimates in a way that is unrelated to the questionnaire. For instance, between 1992 and 1995, the proportion of the unemployed who were teenagers steadily increased from 14.8 percent to 18.2 percent, while the overall unemployment rate steadily declined from 7.5 percent to 5.6 percent. Similarly, the proportion of the unemployed who were Hispanic steadily increased from 13.6 percent to 15.4 percent between 1992 and 1995, though some of this may be due to the increasing proportion of Hispanics in the population (which rose from 8.8 percent to 9.4 percent). Both teenagers and Hispanics tend to be lower skilled workers who historically have been more likely to become unemployed or withdraw from the labor market. It should be noted that, regardless of the
source, an increase in the proportion of the unemployed drawn from groups with less stable labor force histories will influence the MLCA model estimates of accuracy if the change is not accounted for in the modeling.

6. Differential Validity of the Markov Assumptions

In addition to differences in the composition of those classified as on layoff affecting the estimates generated by the MLCA models, differences in the composition of the various labor force groups before and after the redesign could affect the validity of the underlying assumptions of the MLCA models. As the author notes, a key assumption when implementing MLCA models is that an individual’s transition from the second to third month is independent and thus uninfluenced by how the individual was classified in the first month. When estimating MLCA models for individuals’ labor force states this obviously is untrue, and the validity of the assumption will likely differ amongst the various labor force categories. For instance, an individual who is employed in the first month is much more likely to be employed in the third month than is an individual who has never worked. More importantly, an individual cannot be classified as on layoff in either the redesigned or unredesigned questionnaire if he or she has not previously worked. Addition, under the official definition of layoff that was implemented in the redesign, individuals also have to expect to be recalled. This leads to a much tighter relationship between employers and workers across months using the redesigned questionnaire. Given that individuals on layoff under the redesign are much more likely to be recalled and thus employed than under the unredesigned questionnaire, the likelihood of an individual’s labor force status in the third month depending on their initial labor force status in the first month is much higher. Consequently, not only is it likely that the Markov assumptions are often violated in labor force studies; it is much more likely that the Markov assumptions are violated after the redesign. This differential violation of the model’s assumptions could be fundamentally influencing the author’s results.

7. Conclusion

In summary, although the author believes that he identified a problem that was introduced into the CPS with the 1994 redesign, the supposed increase in misclassification of those on layoff in reality reflects the greater precision of the survey questions. Rather than identify a true error, we believe the author may have failed to recognize that the composition of the groups identified as on layoff before and after the redesign were different due to both intentional changes (such as the definition of on layoff being built into the questionnaire or improved quality of matches obtained because of computerization of the survey) and to uncontrolled changes such as developments in the overall economy. Finally, we would like to see further work in this area which combines the MLCA modeling approach along with a careful consideration of the economic concepts being measured, the time periods being examined and the assumptions being made. We believe this could lead to a more accurate understanding of the effects of the 1994 CPS redesign, and more useful application of the MLCA modeling approach in general.

Acknowledgements

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics. The authors would like to thank Sharon Cohany, U.S. Bureau of Labor Statistics, for helpful commentary on this discussion.

References


Comment

Clyde Tucker

1. Introduction

I first would like to congratulate Paul Biemer for offering an innovative approach to the study of measurement error in surveys. Although he chose to illustrate his approach with the employment series in the Current Population Survey (CPS), the method can be applied to many surveys. My comments largely will be conceptual in nature, but I will supplement these comments with examples from the same data that Biemer analyzed.

Using Markov Latent Class Analysis (MLCA), the Biemer paper relies on an evaluation of the consistency over time of respondents’ answers to the questions in the employment series. The increase in inconsistency found in the new series as compared to the old one, after controlling for self versus proxy reports, may serve as an indicator of one type of measurement error in the assignment of labor force status. Presumably, this error is the result of the failure of the new questions (at least, compared to the old ones) to collect the correct information for classifying an individual into the right labor force category. Thus, the error can be attributed to poor question design. Because the analysis indicates that the errors tend to be in one direction more than in the other – the misclassification of truly unemployed individuals into a different category – some might interpret the result to be a bias in the unemployment rate.

I will argue that not only has bias not been introduced but also that the new series, while certainly not perfect, reduces error, providing a more accurate picture of the employment situation. It does this by taking into account the economic realities of today in a way that the old series did not. This is accomplished by not only better question wording but also by the inclusion of follow-up questions and probes that capture more detailed information for determining a respondent’s true employment status. The use of follow-up questions and probes is facilitated by the introduction of a computerized survey instrument. As a result of these innovations, I believe that the new employment series reduces the amount of specification error that existed with the old series. By specification error, I mean the error arising from using questions that do not measure what they are intended to measure. I also will explain why I do not believe that Biemer’s method is appropriate for use in this particular case.

2. Recognition of the Need for a New Employment Series

The last major revision of the CPS prior to 1994 took place in 1967. In the ensuing years, the labor market underwent a great transformation. The number of women in the labor force dramatically increased. The number of part-time jobs and multiple job holdings escalated. The relationship between the worker and the employer became more tenuous. Startling technological developments changed the way Americans did work and resulted in the creation of new types of jobs requiring new kinds of skills. Perhaps most importantly, the economy gradually became more service oriented and less manufacturing oriented.

Just one result of these developments that needed to be taken into account in the CPS was the change in the accepted meaning of “layoff” as so ably described by Miller and Polivka (2004), but there were others, as enumerated by Bregger and Dippo (1993). Better information was needed about discouraged workers (those who have given up looking for work), multiple jobholders, marginal workers (e.g., unpaid workers in a family business), and job-changing patterns. In addition, during the 1970s and 1980s, concern mounted about the various types of nonsampling errors that could be affecting CPS estimates as well as about respondent burden and its detrimental effect on data quality.

Until the 1980s, the technology to tackle these problems was not available. However, as Bregger and Dippo (pages 4–5) note, things began to change:

“…in the early 1980s, the introduction of two new survey methodologies provided the means for understanding and reducing measurement error. These included the application of behavioral science methods and theory – more commonly referred to as the cognitive aspects of survey methodology – and computer-assisted interviewing. It is through the blending of these two methodologies that a new collection procedure, which focuses on reducing measurement error, was made possible.”

Cognitive methods (including focus groups and in-depth interviewing) made it possible to develop questions that

---

could accurately measure the more complex economic behaviors that the times required. Furthermore, these techniques were able to uncover problems in the existing labor force series (see Polivka and Rothgeb 1993). The accurate measurement of the more complex behaviors also required a more complicated survey instrument. One so complicated that interviewers, left to their own devices, would have difficulty navigating. This is where computer-assisted interviewing played an important role. With a computerized survey instrument, interviewers could easily navigate through the complex skip patterns necessary to obtain answers to questions for measuring the wide variety of economic behaviors of interest.

3. Consideration of Nonsampling Errors in Both the Old and New CPS Employment Series

Let me begin this section by detailing my reasons why MLCA is not an effective tool for evaluating the new CPS design relative to the old one. MLCA can be a good method for detecting measurement error within a constant series of questions by looking for inconsistencies in response over several administrations to the same respondent. In the case of the CPS, the method might be appropriate, given a careful examination of a well-chosen set of diagnostics, for examining problems in the old employment series and the new employment series independently of one another. However, let me add a caveat here about examining inconsistencies even within the same employment series. Labor force status, in itself, is inherently inconsistent over time. While the employed and not-in-the-labor-force (NILF) categories are relatively stable, the unemployed category is not. Those in that category are trying to get out. Controlling for seasonal effects by looking at March-May of either 1993 or 1994, it turns out that, on average, almost 90% of those in the employed and NILF categories did not move from one month to the next. On the other hand, over half of those in the unemployed category did. Thus, the unemployed are a particularly difficult group for MLCA to handle.

As for comparing the two series, the use of MLCA is problematic because the two series were designed to measure different things. There were some significant changes made in the employment series in the hopes of reducing specification error. Although I do not want to dwell on the measurement of layoff (Miller and Polivka have covered this topic well), I do want to use it as a case in point for explaining why the comparison of the old and new instrument is a difficult one to make. Apart from what Miller and Polivka have said, I have my own reasons for doubting Biemer’s conclusions.

The changes in the layoff questions were designed to reduce the specification error discovered in qualitative research on the meaning of “layoff,” as alluded to by Miller and Polivka. In the attempt to eliminate specification error, two additional questions were added. One asked whether a date for recall had been given, and the other inquired about the possibility of returning to the job within the next 6 months. Only those who were given a recall date or expected to return to work within the 6-month period were classified as truly “on layoff.”

Clearly, this altered the characteristics of the group classified as unemployed as a result of layoff as well as those asked the remaining questions in the employment series, but I believe there also were more subtle reasons why inconsistencies in respondents’ answers could have increased and still not have contributed to measurement error to the extent argued by Biemer. In the first place, respondents had to answer more questions, which would have increased the probability that at least one false inconsistency would arise from one month to another. This might add to measurement error compared to the old series, but specification error, considered to be the greater problem, still would be reduced. Furthermore, false inconsistencies arising from these questions should be minimized for two reasons. These questions are much more specific than the single layoff question in the old series, and they had been well tested (Esposito, Campanelli, Rothgeb and Polivka 1991). Moreover, given that more specific questions were asked, there would be an increased chance that true change had taken place in the state of at least one of them in the intervening month. Finally, and of greatest interest to me, is the fact that these questions attempt to capture information on relatively nuanced changes. For instance, a respondent may have changed his or her mind about the possibility of being recalled in the next 6 months based on little concrete information. With the uncertainties in today’s job market, it would be difficult to say that the respondent had given the wrong answer.

I now want to address Biemer’s concerns about the initial question in the new employment series asking about whether any work was done last week for “either pay or profit.” His results indicate that this question may be contributing to the amount of error he finds in both the “layoff” and “looking” series. The change in this question (as well as the addition of a question on the existence of a family-owned business or farm) was prompted by the concern that the old questions were not stated broadly enough, so that marginal workers, especially those working for profit at home, were not being classified as working. For example, the Parallel Survey showed the percentage of part-time workers in the new CPS was 1.098 times larger than in the old CPS, and, coincidently, the employment to
population ratio for women 65 and older also increased by about the same amount (Polivka and Miller 1998). The same is true when comparing 1993 to 1994. It stands to reason that the increased precision in the identification of these marginal workers, who are more likely to be inconsistent in their answers from month to month than other workers, might be mistaken for measurement error. The fact is the more narrow “what were you doing last week” question could lead these respondents to consistently, but inaccurately, report they were unemployed.

Finally, let me turn to the other section of the employment series in which Biemer found a problem – the “looking for work” questions. One important change in this series involved clarifying the differences in “active” and “passive” job search in order to reduce misclassification rates in these categories. Studies conducted in the 1980s found that interviewers were confused about what constituted an active (versus a passive) job search (Polivka and Rothgeb 1993). In the redesigned questionnaire, interviewers were given an explicit list of both active and passive job search methods. Comparisons of the results of the old and new questions are complicated by the fact that different subpopulations were asked these questions in the two series. Those finally defined as looking (and, thus, considered unemployed) in the two different employment series could have arrived there in quite different ways. Half of those considered looking in 1993 received that designation by volunteering they were looking in the first question (“What were you doing most of last week?”); none of those who were looking in 1994 followed that path. Those retired and 50 or older in 1994 never got the chance to say they were looking. In 1993, none of those who said they were on layoff were asked the looking question, so they had no chance to be classified as NILF in a given month. Then there were the two different levels of information given to the interviewers for coding active and passive methods. One difference uncovered in an analysis of the two groups from 1993 and 1994 was that a higher proportion of those looking in 1994 were women compared to 1993 (45.4% vs. 41.2%). Referring to the above discussion on the first employment question, increases in the inconsistency in reports to the looking questions could be the result of capturing more marginal workers using the revised employment series. Sometimes these individuals would be looking and sometimes not.

4. Conclusions

Paul Biemer has made a bold attempt to investigate the error structure in the CPS employment series; however, his findings do not take into account the reasons for the revised questions. Taking these into account would help explain the month-to-month inconsistencies that he found. Not only might these inconsistencies be real, but they could provide evidence of a reduction in specification error. For instance, controls other than for self/proxy could be included in the model to take into account some of the changes in methodology, and measurement error within more limited subpopulations. More exploration of the utility of MLCA with inherently inconsistent classifications also should be undertaken.

Acknowledgements

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics. The author would like to thank Steve Miller, Anne Polivka, and John Dixon for their assistance on this discussion.

References


Response from the Author

Paul P. Biemer

1

1. Introduction

My sincere thanks to all four discussants for their thoughtful, thorough and constructive comments. They have added considerably to our understanding of the complex issues surrounding Markov Latent Class Analysis (MLCA) and the Current Population Survey (CPS) labor statistics. All four discussants raise a number of important issues that I will try to address to the extent I can. Some issues will require more work and deserve much greater consideration than is possible here. More complete responses to those issues will have to await the results of future research.

Considering all the comments collectively, there seems to be agreement that Markov latent class analysis has considerable potential as a tool for evaluating and exploring the sources of measurement error in the CPS. However, there is some skepticism that it has identified real problems in the CPS questionnaire. Dr. Vermunt, who is also the author of the software I used for this analysis (viz., EM), provides a number of valuable suggestions for improving the models and investigating the validity of the model assumptions. The three other reviewers (Drs. Miller, Polivka, and Tucker) are quite familiar with the CPS since they are employed by the federal agency that sponsors the survey where they played important roles in the 1994 redesign. Their comments remonstrate the various ways in which the MLCA model assumptions could be violated for these data. In addition, they contain valuable information regarding details of the CPS (both pre- and post-redesign) and the construction of the CPS labor force variable. The comments and suggestions of all the discussants should be carefully considered by labor force economists and statisticians who are conducting research in the area of employment measurement error, particularly those using MLCA.

Jeroen Vermunt’s Comments

I first address the comments of Dr. Vermunt and then the comments of the other three reviewers. I share Dr. Vermunt’s concern that the ICE assumption may not hold for these data. As he points out, if respondents misunderstand the labor force questions in the same way from one month to the next, they may make the same errors each month creating correlated errors across the months. As an example, a person who is truly in the UEM category at both Times 1 and 2 may be more likely to be misclassified at Time 2 if they were also misclassified at Time 1. This can be stated probabilistically as

\[
\rho = \frac{P(B \neq 2 | A = 2 \text{ and } X = Y = 2)}{P(B \neq 2 | A = 2 \text{ and } X = Y = 2)} - 1 > 0. \tag{1}
\]

The numerator probability of the quantity \(\rho\) is the probability that the Time 2 classification \((B)\) is in error given the Time 1 classification \((A)\) is also in error and the true classification at both time points is UEM. The denominator probability is similar except for the condition that no error is made at Time 1 \(\text{i.e.}, A = 2\). Under the ICE assumption, \(\rho = 0\). Therefore, if the \(\rho > 0\) (which is the likely direction of the correlated error), the ICE assumption is violated. Dr. Vermunt suggests a simulation study be conducted to study the sensitivity of the estimated classification errors to violations of this assumption. Of course, determining the extent to which the ICE assumption fails for the CPS data is not possible via simulation. Nevertheless, it is still useful for assessing the potential for correlated error to bias the MLCA classification error estimates.

Following his suggestion, I conducted a small simulation study to gain some insight as to the consequences \(\rho > 0\) for MLCA using CPS data. A sequence of artificial populations was generated using parameters consistent with those for the CPS (see for example, Table 1 in the main paper) except that \(\rho\) was increased in small increments from 0 to its empirical maximum – \(\text{i.e.}, \rho\), the largest value of \(\rho\) that is feasible without violating the other model assumptions. Maintaining the other model assumptions in the analysis is necessary so that the consequences of violating just the ICE assumption can be isolated.

The largest feasible value of \(\rho\) was determined empirically to be 0.7. At this value of \(\rho\), the MLCA estimate of the probability a correct classification of UEM went from 79% to 85% and the misclassification error rate dropped from 21% to 15%. For mild departures from the ICE assumption,
say $0 < \rho < 0.3$, the error rates changed by less than 3 percentage points. These results illustrate that if the ICE assumption fails to hold due to positive between interview correlations, the error rates estimated by MLCA will be somewhat underestimated. However, mild departures from the ICE assumption should have little effect on the classification error probabilities for these data. A similar analysis was conducted for the two other labor force categories (i.e., EMP and NLF) but the change in the classification error estimates was negligible. This result was anticipated due to the relatively small error rates for these categories.

The results suggest that mild departures from the ICE assumption should have little or no effect the conclusions of the analysis. Extreme departures might affect the conclusions in the unlikely event that errors are highly correlated for original questionnaire and essentially uncorrelated for the revised questionnaire. Under that scenario, the original questionnaire would appear to have smaller UEM classification error than the revised questionnaire. However, there is no practical reason to expect this condition to hold since both questionnaires present questions that respondents may misunderstand consistently across interviews.

Although these simulation results, as well as those in Biemer and Bushery (2001) for investigating the consequences of violations of the Markov assumption, are quite useful for studying the sensitivity of the estimates to violations of the MLCA model assumptions, they provide no direct evidence of the validity of the MLCA estimates. Biemer and Bushery (2001) illustrate how the (empirical) validity of latent class estimates can be established using external data and alternative approaches for estimating classification error. A similar analysis based upon test-retest reinterview data will be provided in the sequel.

For the purpose of identifying potential areas where the CPS questionnaire can be improved, it is not essential to establish unequivocally that the MLCA model assumptions hold since model validity is of secondary importance. Instead, the primary issue for questionnaire evaluation work is whether the method of analysis used is successful at identifying questions that have large measurement errors and are in need of revision. In other words, the validity of the model is established by its ability to find important flaws in the questionnaire. Determining whether there truly is error in the UEM classification as suggested by MLCA requires an evaluation using other methods such as cognitive laboratory research. Cognitive interviews could be used to investigate encoding, comprehension, recall, and/or social desirability issues that generate errors in the responses to the UEM questions. If these investigations uncover important problems in questions, then the utility of MLCA for identifying flawed questions will be supported even though the validity of the MLCA modeling assumptions may never be known.

Dr. Vermunt’s other suggestions on ways the modeling framework could be improved are quite reasonable and I hope to investigate them further in the future. However, the current software for fitting MLCA models is somewhat limited and the estimation of complex models such as those he suggests may not be feasible. He also notes that problems can arise when fitting large models with the EM algorithm. As an example, initially we attempted to use the proxy/self-response variable as a time-varying covariate in the MLCA models, but encountered problems in the estimation process such as “division by 0” errors and persistent convergence to local maxima. We ultimately had to abandon the approach in favor of the single, time invariant proxy/self grouping variable used in the current analysis. As new and more general software becomes available, the options for MLCA with time varying covariates as well as other model enhancements mentioned by Dr. Vermunt will be feasible.

Comments of the BLS Discussants

I will address the comments of Drs. Miller and Polivka and those of Dr. Tucker together since the reviewers are from the same agency (BLS) and their comments raise similar concerns about the analysis. The following five points summarized their main concerns:

1. The modifications introduced in the new questionnaire capture more transitions than the old questionnaire. MLCA wrongly interprets these as errors when in fact they are not error.

2. Respondents may change their minds from month to month about whether their employers truly indicated that they might be recalled to work. These changes should not be classified as a response error.

3. The Markov assumption does not hold in labor force studies and it is violated to an even greater extent after the redesign than before the redesign. This differential violation of the model’s assumptions could be fundamentally influencing the MLCA results.

4. The differences in the estimates of LAYOFF classification error before and after the redesign are due to the composition of the groups comprising this category. This composition changed after the redesign in a manner that was desired and intended by those who redesigned the questionnaire.
5. The increased inconsistency in reports to the LOOKING questions for the revised questions could be explained by more marginal workers being identified using the revised questions. Sometimes these individuals would truly be looking for work and sometimes not. MLCA misinterprets these ostensibly random changes as response error when they are not.

Point 1 describes an issue that should not pose any difficulties for MLCA. The MLCA model assumes that each individual occupies a true labor force state which may change from month to month. No assumption is made that the transition probabilities are the same for both questionnaires. The true initial labor force probabilities as well as the month-to-month transition probabilities are estimated independently for each questionnaire. In fact, although not discussed in main paper, the model estimates of the true exit probabilities for LOOKING and LAYOFF are in fact greater for the revised questionnaire than for the original questionnaire. Thus, a greater number of flows from one labor category to another for the revised questionnaire does not necessarily bias the estimates of classification error for that category in either direction.

Point 2 suggests that whether an individual is truly on layoff depends upon that individual’s opinion about whether he or she was given an indication of possibly being recalled. However, in this not how the revised questionnaire defines the concept. An individual’s true layoff status depends upon whether or not the employer truly provided an indication of being recalled. Although the respondent’s opinion about what the employer indicated may change from month to month, the true layoff status does not change according to the respondent’s opinion. Flows in and out of the LAYOFF category due to the respondent’s opinion should be interpreted as error by the model.

Points 3, 4, and 5 could be made for any analysis employing MLCA. They essentially concern the potential bias in the MLCA estimates when month-to-month transitions do not behave according to the MLCA model and consequently real changes are misinterpreted as classification errors. As the reviewers note, there are at least three ways this can occur:

a) the Markov assumption does not hold (point 3),

b) there is unobserved or unexplained heterogeneity in the population (point 4), and

c) employment-related behaviors for two consecutive months are not correlated for some persons; thus, for those persons, past month status does not predict the current month’s status (point 5 as well as a point made by Dr. Vermunt).

The implications of (a) were considered in a simulation analysis in Biemer and Bushery (2001). Their results suggest that, for the CPS data, the estimates of classification error are quite robust to violations of the Markov assumption. It is unlikely, then, that non-Markov transitions explain the findings of higher classification error for the revised questionnaire. Still, additional research is needed to more thoroughly understand the implications of non-Markov transitions for our results.

For (b), it is quite possible for MLCA estimates to be biased when the compositions of the unemployed populations are substantially different under the original and revised questionnaires and those differences are not explained by the grouping variables used in the model. Likewise (c) may be regarded as a special case of (b). For (c), the transition probabilities for some population subgroup are uncorrelated with the prior month’s employment status; instead it is correlated with other unobserved variables. In Jeroen Vermunt’s coffee drinker example, the unobserved variable is the availability of a specific brand of coffee at the market. At this stage of the research, we have not conducted simulation studies to quantify the effects of unobserved heterogeneity on the estimates, but this possibility will be examined in future work.

However, this issue as well as the general plausibility of the MLCA estimates can be investigated to some extent by comparing the MLCA estimates with independent estimates from an estimation approach that is not affected by (a) through (c). If the findings from the alternative analysis are consistent with the MLCA findings, the MLCA findings gain credibility. As an example, test-retest reliability for the CPS employment classifications can be estimated both pre- and post-redesign using the CPS reinterview data (see for example Biemer and Forsman 1992 for a description of CPS reinterview program and these data). The validity of the estimates of test-retest reliability does not depend upon the Markov assumption or group homogeneity assumption; the ICE assumption, however, is still relevant for reliability estimation.

Table 1 shows estimates of Cohen’s kappa measure of reliability for three time periods: 1992–1993, 1995–1997, and 2002–2003. As shown in the table, the reliability of the CPS classifications of unemployment dropped after the redesign from about 68% to 65%. The most recent estimates of kappa indicate reliability has dropped to below 60%. These results are consistent with the results from the MLCA that classification error in the CPS unemployment statistics has worsened after the redesign. It is possible that the reliability estimates in Table 1 are biased since they also rely on the validity of the ICE assumption. But as discussed previously, in order to the results in the table to be explained by the failure of the ICE assumption, the ICE assumption
would have to hold for the revised questions but not for the original questions. That condition is very unlikely to occur.

<table>
<thead>
<tr>
<th>Year</th>
<th>n</th>
<th>Cohen’s κ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992 – 1993</td>
<td>28,063</td>
<td>67.8</td>
</tr>
<tr>
<td>1995 – 1997</td>
<td>22,429</td>
<td>64.6</td>
</tr>
<tr>
<td>2002 – 2003</td>
<td>19,205</td>
<td>58.8</td>
</tr>
</tbody>
</table>

1 From Biemer and Bushery 2000.
3 Personal communication with Bac Tran at the U.S. Census Bureau.

Given the evidence presented here and in the main paper, it seems reasonable to consider the possibility that CPS unemployment classification error increased after the redesign. The next step is to conduct additional research to evaluate these findings and explore the possible causes for the error. Rather than to focus on the validity of the MLCA or test-retest reinterview models, the focus of the future research should be the revised CPS questions, particularly those used in the LAYOFF classification.

I have already mentioned the possibility of using cognitive interviews to investigating the problems in the response process associated with the revised questions. As an example, one question identified in the MLCA as being potentially flawed is: “Have you been given any indication that you will be recalled to work within the next 6 months?” Some of the issues that could be investigated in the cognitive laboratory for this question include:

- How well do unemployed subjects understand the meanings of terms such as “any indication” and “recalled?”
- Do subjects who were recently separated from employment have difficulty remembering what their employers said about being recalled when they were terminated?
- An employer may say, “If business improves, we may call you.” Do respondents answer the question correctly in this situation?
- Do respondents who initially respond that they will be recalled later change their responses to this question as the months pass by and they have not been recalled?

### Specification Error and Measurement Error

Finally, I will address an important issue raised by Dr. Tucker regarding specification error, measurement error and their net effects. As Dr. Tucker explains, the original questionnaire suffered from specification error bias caused by measuring the wrong concept. The revisions to the labor force questions introduced in 1994 were designed to eliminate the specification error bias by refining the concepts of employment and unemployment and modifying the survey questions to reflect these refinements. These modifications, while reducing specification error, added more complexity to the survey questions which could have increased the measurement error bias in the labor force estimates. Dr. Tucker suggests that while this may be the case, the measurement bias in the new employment series may be less than the combination of specification bias and measurement bias in the old series. To determine whether this could be true, the specification error bias ($B_s$) and measurement error bias ($B_m$) were separately estimated using the MLCA estimates provided in the paper as described below.

Let $p$ denote the CPS estimate of UEM and let $P$ denote the expectation of $p$ with respect to sampling and measurement error distributions. Let $\pi$ denote the true value of the characteristic under the definitions of UEM implied by the specific questionnaire (i.e., without regard to possible specification error). Therefore, $\pi = P - B_m$, i.e., the value of $P$ in the absence of measurement error bias.

As noted above, specification error bias is the bias in $P$ due to a wrong concept or definition of unemployment implied by the questions and/or labor force classification process. For the revised questionnaire design, we assume that the specification error in $p$ is 0 since it will be regarded as the gold standard for estimating the specification error bias in the original questionnaire.

Let $\pi_{\text{old}}$ and $\pi_{\text{new}}$ denote the $\pi$-parameter for the original and revised questionnaires, respectively. Then the specification error bias in the pre–1994 estimates of the unemployment rate is

$$B_s = \pi_{\text{old}} - \pi_{\text{new}}.$$  \hfill (2)

For each questionnaire, the estimate of $P$ is $p$, the weighted estimate from the CPS. The estimate of $\pi$ is obtained by correcting $p$ for classification error bias using the response probabilities from the MLCA. Let $p' = (p_1, p_2, p_3)'$ where $p_1, p_2, p_3$ denote the estimates of the proportions in EMP, UEM, and NLF, respectively. Let $o_{ij}$ be the probability that an observation that truly belongs to the $j^{th}$ category is assigned to the $i^{th}$ category and let $\omega$ denote the true proportion in the population in the $i^{th}$ category. Then

$$E(p) = \Omega \pi$$  \hfill (3)

where $\pi = (\pi_1, \pi_2, \pi_3)'$ and $\Omega = [o_{ij}]$ is the $3 \times 3$ matrix with elements $o_{ij}$. It follows that an estimator of $\pi$ is

$$\hat{\pi} = (\hat{\Omega})^{-1} p.$$  \hfill (4)
where \( \hat{\Omega} \) is a MLCA estimate of \( \Omega \). For each questionnaire, \( \hat{\Omega} \) was estimated by the average of the 10 MLCA estimates (January–March through October–December) using the 1993 CPS for the original questionnaire and 1993 Parallel Survey for the revised questionnaire.

Table 2 shows the results of this analysis. For UEM, \( p = 6.38 \) for the original and 6.98 for the revised questionnaire. If the unemployment rates are corrected for measurement bias using (4), unemployment rate increases to 7.09 percent for the original questionnaire and 8.03 percent for the revised questionnaire. Thus, an estimate of the measurement bias for the original survey is \( 6.38 – 7.09 = –0.71 \) and for the revised survey is \( 6.98 – 8.03 = –1.05 \). Note that the measurement biases are negative for both the original and revised questionnaires, indicating that UEM as well is underestimated by both questionnaire versions.

For the revised questionnaire, the specification bias is assumed to be 0. For the original questionnaire, it is estimated by the difference \( 7.09 – 8.03 = –0.94 \) percent. An estimate of the net bias, \( B_T = B_M + B_S \), is \( –0.71 + (–0.94) = –1.65 \) percent for the old series compared with \( –1.05 + 0 = –1.05 \) percent for the new series. Thus, while it is subject to greater measurement error bias, the new series has smaller estimated net bias assuming \( B_S = 0 \).

Several limitations of these results should be mentioned. First, as noted in the main paper, the estimates for revised questionnaire from the Parallel Survey may not be representative of the revised CPS series. Second, the analysis assumes that the revised questionnaire is the gold standard for estimating the specification error bias in the original questionnaire. This assumption could also be challenged. Finally, no standard errors were provided for the estimates in Table 2 and the hypothesis of smaller overall bias in the revised question was not formally tested. Despite these limitations, the results suggest the possibility that the new unemployment series could have substantially lower net bias than the old series.

### Table 2

Comparison of Original and Revised Questionnaire Biases for the CPS Unemployment Rate Based Upon Estimates from the 1993 CPS and the Parallel Survey

<table>
<thead>
<tr>
<th></th>
<th>( p )</th>
<th>( \pi )</th>
<th>( B_M )</th>
<th>( B_S )</th>
<th>( B_T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993 CPS</td>
<td>6.38</td>
<td>7.09</td>
<td>–0.71</td>
<td>–0.94</td>
<td>–1.65</td>
</tr>
<tr>
<td>Parallel Survey</td>
<td>6.98</td>
<td>8.03</td>
<td>–1.05</td>
<td>0(^1)</td>
<td>–1.05</td>
</tr>
</tbody>
</table>

\(^1\)Note: Specification error bias is assumed to be 0 for the revised questions.

### References

Biemer, P., and Bushery, J. (2001). Application of markov latent class analysis to the CPS. *Survey Methodology, 26*, 2, 136-152.