Survey Research Methods (2011)
Vol.5, No.2, pp. 39-51
ISSN 1864-3361
http://www.surveymethods.org

Estimating underreporting of consumer expenditures using Markov latent class analysis

Clyde Tucker
Bureau of Labor Statistics

Paul P. Biemer
RTI International

Brian Meekins
Bureau of Labor Statistics

This paper examines reporting in specific consumer item categories (or commodities) and estimates expenditure underreporting due to survey respondents who erroneously report no expenditure in a category. Our approach for estimating underreporting errors is a two-step process. In the first step, a Markov latent class analysis is performed to estimate the proportion of consumers in various subpopulations who fail to report their actual expenditure in a particular commodity. Once this proportion is estimated, the dollar value of the missing expenditure is estimated using the mean expenditure of those in that subpopulation that did report an expenditure. Finally, the estimates are evaluated and discussed in light of external data on expenditure underreporting.

Keywords: underreporting error, consumer price index, missing data, survey evaluation

1 Introduction

This paper examines the quality of reports for specific categories of consumer expenditures (referred to as *commodities*) in the Consumer Expenditure Survey's (CE's) Interview Survey (CEIS) component. It also attempts to quantify the magnitude of underreporting for respondents who claim no expenditure for a commodity (i.e., respondents reporting zero expenditure who, in fact, should report a positive expenditure). The accuracy of nonzero expenditure reports is also an important problem, but it will not be considered here.

We use a two-step approach to estimate the magnitude of erroneous zero underreporting errors. In the first step, a Markov latent class analysis (MLCA) is performed to estimate the proportion of consumers, in a specified subpopulation, who fail to report an actual expenditure in a particular commodity (so-called *false nonpurchaser*). Once this proportion is estimated, the second step estimates the dollar value of the missing expenditure from the expenditures of those in the same subpopulation who did report an expenditure for the commodity (so-called *purchasers*). As we shall see, consumers reporting a fictitious expenditure (so-called *false purchasers*) are extremely rare and, therefore, the focus in this paper is primarily on false nonpurchasers.

MLCA, first proposed by Wiggins (1973), extends the ideas of latent class analysis (LCA) to panel survey data. LCA, developed by Paul Lazarsfeld (1950), treats the true value of a questionnaire item (such as whether a particular type of expenditure was incurred during the time period) as an unobservable (latent) variable. It specifies a model for this latent variable, taking into account the interrelationships between observed indicator variables and the grouping (subpopulation) variables. The theoretical development of this model originated with Lazarsfeld and Henry (1968) and Goodman (1974).

The paper begins with an introduction to the CEIS, which is sponsored by the U.S. Bureau of Labor Statistics (BLS) and conducted by the U.S. Census Bureau. The paper then discusses the mathematics of the modelling of the latent variable, as well as previous research by the authors in this area. The mathematics for producing an estimate of a missing expenditure is given, and the results of the analysis follow. A final section discusses the results and considers additional avenues for research.

2 Background

In the CEIS, respondents are asked about the amounts and descriptions of the expenditures they incurred in the last quarter of the year for a long list of household commodities. The detailed questions are asked for a commodity only if the respondent responds positively to a screening question that essentially determines whether any purchases of the item have been made over the past 3 months. If the response to the screening item is negative, then detailed questions are skipped.

Because of the large number of items in the survey and the extensive information requested for some purchases, interviews can last from 1 to 2 hours or longer, and the burden on respondents can be considerable. Although the potential for underreporting expenditures as a result of false negative responses to the screening questions is apparent, few studies have formally investigated the error in CEIS screening questions. The primary reason is that these types of errors are difficult to assess by conventional statistical techniques.

Contact information: Clyde Tucker, Bureau of Labor Statistics, e-mail: nctucker@cox.net

Traditional methods for evaluating survey error require the use of *gold-standard* measurements; i.e., essentially error-free measurements that may be regarded as the truth for purposes of estimating measurement error. In the CEIS, such gold-standard measurements are very difficult to obtain. Every 5 years, comparisons between aggregated expenditures by commodity from the CEIS are made to Personal Consumption Estimates (PCEs)[1] calculated from administrative records by the Bureau of Economic Analysis (BEA). Although these comparisons often indicate that there is considerable underreporting of expenditures in the CEIS, they do not provide information on how the underreporting is related to the characteristics of individual respondents. Furthermore, the comparisons are available infrequently and are not useful for timely quality-improvement research in the CEIS program.

The option of intensive interviewing methods, which could provide some information at the consumer level for evaluating the underreporting, is also not viable. The sample sizes required would have to be sufficiently large to provide precise estimates of the error rates, so the costs of the study would be substantial. Even if such a study were conducted, the validity of its results would be problematic because the excessive respondent burden is likely to lead to a great deal of nonresponse. In addition, to be accurate, the collection of expenditure information would have to rely on purchase receipts and household billing records which are not routinely retained in all households for many expenditures. Experimental methods could provide some insights, but again the sample size would be necessarily small and unlikely to be representative.

In lieu of a gold standard, MLCA provides information on underreporting that can be linked to individual respondents. It requires neither greater respondent burden nor expensive additional data collection. MLCA assumes an error model for the available measurements and estimates the parameters of the error model using maximum likelihood. Thus, the validity of the MLCA estimates hinges on the ability of the model to accurately represent the error-generating process. MLCA requires at least three panel measurements of the same item as a condition of estimability of the model parameters. The MLCA model then specifies parameters for both the period-to-period changes in the status of the item as well as the measurement error associated with expenditure reports.

This paper continues the analysis of the CEIS along the lines investigated in a number of previous papers. Biemer (2000) applied the MLCA methodology to the CEIS to determine whether useful information on the magnitudes and correlates of screening question reporting error can be extracted directly from the CEIS panel data. Biemer and Tucker (2001) extended the earlier analysis using data from four consecutive quarters of the CEIS by considering consumer units (CUs) that were interviewed for four consecutive quarters during the period beginning with the first quarter of 1996 and ending with the last quarter of 1998. This allowed the authors to consider a wider range of models, including second-order Markov models. First-order Markov models assume that a purchase or nonpurchase at quarter q is affected only by quarter $q-1$ purchases or nonpurchases. A second-order Markov model assumes that the two previous quarters, i.e., $q-1$ and $q-2$, affect purchasing behavior at quarter q. Their analysis provided evidence of second-order Markov effects and recommended that second-order terms be included in the models.

Tucker, Biemer, and Vermunt (2002) fitted very similar models to these data using both unweighted and weighted data. Comparing the estimates from the two approaches, they reported very few differences in the error parameter estimates and recommended that the sampling weights be ignored in the analysis when the analysis is focused on classification error. Thus, following their recommendation, unweighted analyses will be used for the present paper. In addition, Tucker, et al. (2002) conducted a thorough examination of the explanatory variables considered in the previous studies and identified a reduced set variables that are most predictive of misclassification. These variables will form the starting point for model selection in our analysis.

3 Data Sets and Models

This section describes the data sets that will be analyzed in the study and the essential components of the MLCA methodology. The data consist of 8,817 interviews collected in 3 years of the CEIS: 1996, 1997, and 1998. Each of these surveys was designed to collect information on approximately 95 percent of the total household expenditure. In the CEIS, a CU is defined as the members of a household who are related and/or who pool their incomes to make joint expenditure decisions. The CEIS interviews CUs once every 3 months for five consecutive quarters to obtain the expenditures for 12 consecutive months. The initial interview for a CU is used as a bounding interview to collect expenditures prior to the reference period to control telescoping, and these data are not used in the estimation. The survey is designed to collect data on major items of expense that respondents can be expected to recall for 3 months. New panels are initiated every quarter of the year so that, in each quarter, 20 percent of CUs are being interviewed for the first time. Only CUs that completed all five interviews are used in this analysis.

MLCA models require that both the latent variables and their indicators be categorical variables. For example, this analysis considers the screening questions in the CEIS for

[1] The PCEs are estimates of expenditures made directly by households that are produced by the BEA. The PCE also includes expenditures made on behalf of households by nonprofit organizations and government programs, such as Medicare and Medicaid. Unlike the CE, the PCE excludes person-to-person transactions. The data for the PCE are collected from a number of surveys, most significantly the Economic Census, which is conducted by the Census Bureau every 5 years. To produce the PCE, the BEA collects receipts from establishments indicating the value of commodities (services and durables), as well as data to estimate taxes, transportation cost, and trade margins. This value is apportioned to the various sectors to which the commodity is sold: government, exporters, and industry, with the residual allocated to the household sector and the PCE.

which the indicator variable is a dichotomous response taking the value 1 if the CU reports a purchase for a particular commodity for the quarter and 2 if not. The corresponding latent variable is also a dichotomous variable representing the true, but unknown, purchasing status for the CU for the quarter. It takes the value 1 if the CU truly purchased the commodity and 2 if not.

Let the CEIS target population be divided into L groups or domains and let the variable G denote a categorical grouping variable that is related to expenditure behavior, underreporting behavior, or both. For example, G may be related to the administration of the survey (e.g., interview length, whether household records were used, number of times the CU had been interviewed previously) or may describe CU characteristics (e.g., size, income, age of CU members). Let, $G_i = 1$ if the ith population member is in group 1, $G_i = 2$ if in group 2, and so on. For ease of exposition, a single variable grouping variable will be assumed in developing the model. In practice, G may be a combination of several variables.

Let the subscript combination (g, i) denote the i-th CU in group $G = g$ for $g = 1, \ldots, L$ and $i = 1, \ldots, n_g$, where n_g is the number of CUs in group g. For a particular 3-month interview period under investigation, let the indictor variable $A_{gi} = 1$ if CU (g, i) reports the item was purchased during the first interview period and $A_{gi} = 2$ if CU (g, i) reports the item was not purchased during this period. For analyzing the data from four consecutive interviews, we define B_{gi}, C_{gi}, and D_{gi} analogously for periods 2, 3, and 4 respectively.

Associated with each of the four observed variables is a latent variable for the true quarterly purchase status of the CU for each time period. For four time periods, the latent variables W_{gi}, X_{gi},Y_{gi}, and Z_{gi} are defined corresponding to A_{gi}, B_{gi}, C_{gi}, and D_{gi}, respectively. Thus, $W_{gi} = 1$ if CU (g, i) truly purchased the item in period 1 and $W_{gi} = 2$ if not. X_{gi},Y_{gi}, and Z_{gi} are defined analogously for periods 2, 3, and 4. Thus, W_{gi}, X_{gi},Y_{gi}, and Z_{gi} refer to true purchasers and nonpurchasers whileA_{gi}, B_{gi}, C_{gi}, and D_{gi} refer to observed purchasers and nonpurchasers for the four quarterly interview periods.

For notational convenience, the subscripts (g, i), are dropped but the relationship of the unsubscripted variable to an individual unit within a group is retained. This notation also implies the "group homogeneity" assumption meaning that the model probabilities, to be defined next, are the same for all individuals belong to the same group, g.

Our notion conforms to the conventions of the MLCA literature (see, for example, Biemer (2011) and Vermunt (1997). Let $\pi_{wxyz|g}$ denote the probability $\Pr(W = w, X = x, Y = y, Z = z \mid G = g)$, let $\pi_{w|g}$ denote $\Pr(W = w \mid G = g)$, $\pi_{x|wg}$ denote $\Pr(X = x \mid W = w, G = g)$, and so on. Then, the probability that an individual in group g has purchase status w in time period 1, x in time period 2, y in time period 3, and z in time period 4 is given by the following:

$$\pi_{wxyz|g} = \pi_{w|g}\pi_{x|wg}\pi_{y|wxg}\pi_{z|wxyg} \tag{1}$$

Note that, for dichotomous variables, this expression contains 30 parameters and the model is not identifiable in a Markov latent class modeling context. If the second-order Markov assumption is invoked, the last term of the model can be written as $\pi_{z|xyg}$ and, thus, (1) can be rewritten as

$$\pi_{wxyz|g} = \pi_{w|g}\pi_{x|wg}\pi_{y|wxg}\pi_{z|xyg} \tag{2}$$

The second order Markov assumption therefore reduces the number of parameters by eight and results in an identifiable[2] model (for a discussion of identifiability, see Biemer 2011).

Using an extension of the notation established previously, let $\pi_{a|gw} = P(A = a|G = g, W = w)$ with analogous definitions for $\pi_{b|gx}$, $\pi_{c|gy}$, and $\pi_{d|gz}$. Thus, $\pi_{a=2|g,w=1}$ is the probability of a false negative report for a member of group g (i.e., the CE classified a CU in group g as a nonpurchaser $[A = 2]$ when the true status is purchaser $[W = 1]$ with respect to the commodity of interest). Likewise, $\pi_{a=1|g,w=1}$is the probability that the CE correctly classifies a person in group $g$as a purchaser. The other response probabilities are defined analogously.

We further assume the classification errors for all four indicators are mutually independent across the four time points referred to as the independent classification errors (ICE) assumption. This assumption can be tested to some extent through the addition of certain interaction terms in the model. However, the assumption seems plausible given that interviews are separated by three months and most causes of ICE failure are believed to be a result of memory effects between panel interviews. This means we can write the joint conditional probability of A, B, C, and D as

$$\pi_{abcd|gwxyz} = \pi_{a|gw}\pi_{b|gx}\pi_{c|gy}\pi_{d|gz}. \tag{3}$$

Finally, we assume that the error probabilities are the same across all four time periods. That is,

$$\pi_{a|gw} = \pi_{b|gx} = \pi_{c|gy} = \pi_{d|gz}. \tag{4}$$

Thus, under these equality constraints, the discussion regarding the estimates of misclassification can be confined to period 1 (i.e., $\pi_{a|gw}$) because the estimates for the other three time periods are assumed to be the same.[3]

Under these assumptions, the probability of classifying a CE sample member in cell (g,a,b,c,d) of the GABCD table can be written as

$$\pi_{gabcd} = \sum_w \sum_x \sum_y \sum_z \pi_g \pi_{w|g}\pi_{a|gw}\pi_{x|gw}\pi_{b|gx}\pi_{y|gxw}\pi_{c|gy}\pi_{z|gyx}\pi_{d|gz}, \tag{5}$$

subject to the time homogeneous error probability constraints.

[2] An identifiable model essentially means that there is but one solution to the maximum likelihood estimation (MLE) equations.

[3] Note that it is possible to relax these constraints somewhat and still achieve an identifiable model. For example, we could constrain any two pairs of error probabilities to be equal and allow difference between the pairs (see, e.g., Langeheine and van der Pol 1990); for instance, $\pi_{a|gw} = \pi_{b|gx}$ and $\pi_{c|gy} = \pi_{d|gz}$. These constraints are not explored in this analysis.

Under multinomial sampling, the likelihood function for the GABCD table can be written as

$$\text{P(GABCD)} \propto \prod_g \prod_a \prod_b \prod_c \prod_d (\pi_{gabcd})^{n_{gabcd}}, \qquad (6)$$

where π_{gabcd} is given by (5) and n_{gabcd} is the number of individuals in cell (g,a,b,c,d). 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 EM algorithm (Dempster, Laird and Rubin 1977) for estimating the parameters of this model and conditions for their estimability. These methods are implemented in the ℓEM software (Vermunt 1997) that will be applied to the CE data set.

To see how the EM algorithm works in the present application, consider the simple case in which there is no grouping variable, G. Now there are just 13 parameters, namely, $\pi_{w=1}, \pi_{x=1|w}, w = 1,2, \pi_{y=1|wx}, wx = 11,12,21,22$, $\pi_{z=1|xy}, xy = 11,12,21,22$ and $\pi_{a=1|x}, x = 1,2$. For the initial E-step, starting values are specified for all 13 parameters that may be randomly generated in the open interval (0,1). Then the full WXYZABCD table is estimated from these initial values of the parameters using equations of the form

$$\hat{n}^{(0)}_{wxyzabcd} = n_{abcd}\hat{\pi}^{(0)}_{wxyz|abcd}, \qquad (7)$$

where $\hat{n}^{(0)}_{wxyzabcd}$ is the estimated count for cell (w,x,y,z,a,b,c,d), n_{abcd} is the observed frequency in cell (a,b,c,d) of the ABCD table, the (0) superscript indicates the 0th iteration of the algorithm, the "hat" symbol denotes an estimate, and

$$\hat{\pi}^{(0)}_{wxyz|abcd} = \frac{\hat{\pi}^{(0)}_w \hat{\pi}^{(0)}_{x|w} \hat{\pi}^{(0)}_{y|wx} \hat{\pi}^{(0)}_{z|xy} \hat{\pi}^{(0)}_{a|w} \hat{\pi}^{(0)}_{b|x} \hat{\pi}^{(0)}_{c|y} \hat{\pi}^{(0)}_{d|z}}{\sum_{wxyz} \hat{\pi}^{(0)}_w \hat{\pi}^{(0)}_{x|w} \hat{\pi}^{(0)}_{y|wx} \hat{\pi}^{(0)}_{z|xy} \hat{\pi}^{(0)}_{a|w} \hat{\pi}^{(0)}_{b|x} \hat{\pi}^{(0)}_{c|y} \hat{\pi}^{(0)}_{d|z}} \qquad (8)$$

is the estimate of the conditional probability $\pi_{wxyz|abcd}$ for the initial iteration. This constitutes the 0th iteration of the E-step. Note that, for the 0th iteration, the parameter values on the right hand side of (8) are just random starting values for the parameters. At the initial M-step, new estimates of these parameters are obtained by applying standard maximum likelihood estimation to the WXYZABCD table for the model with likelihood kernel

$$\pi_w \pi_{a|w} \pi_{x|w} \pi_{b|x} \pi_{y|xw} \pi_{c|y} \pi_{z|yx} \pi_{d|z} \qquad (9)$$

and the constraints in (4), while treating W, X, Y and Z as manifest variables. As an example, the new estimate of $\pi_{w=1}$ is simply

$$\hat{\pi}^{(1)}_{w=1} = \frac{\sum_{xyzabcd} \hat{n}^{(0)}_{1xyzabcd}}{n}. \qquad (10)$$

Then in the second E-step, a new WXYZABCD table is created with frequencies $\{\hat{n}^{(1)}_{wxyzabcd}\}$ given by (7) after replacing $\hat{\pi}^{(0)}_{wxyz|abcd}$ with $\hat{\pi}^{(1)}_{wxyz|abcd}$. The second M-step using these data to obtain new estimates using a full data likelihood maximization approach. Thus, the algorithm continues until convergence is attained.

One model of particular interest in this analysis is the so-called mover-stayer model (Blumen, Kogan and McCarthy 1955; Goodman 1961). For this model, a new latent variable is defined to identify a mover or stayer in the population. A "stayer" is defined as an individual whose purchase status has zero probability of changing for the time periods represented in the model. Thus, for all four quarters a respondent is always a purchaser (represented as "1111") or always a nonpurchaser (represented as "0000"). A "mover" is an individual whose purchase status has a positive probability of changing from quarter to quarter, and over the survey period can be represented by a mix of 1s and 0s. As an example, a current subscriber to cable television is a "stayer purchaser" of cable television if for the four quarters considered for the CEIS, they have probability 1 of remaining a cable television subscription purchaser. Similarly, a "stayer nonpurchaser" is a person who does not subscribe to cable television in quarter 1, and has zero probability of ever subscribing for quarters 2, 3, and 4. For the cable/satellite television commodity, we expect a high proportion of the population to be either stayer-purchasers or stayer-nonpurchasers because transitions in and out of purchaser status are rare for this commodity over four quarters.

However, for the shoes commodity, the proportion of movers is likely to be higher than stayers because shoe purchases typically happen more frequently (if sporadically) during a 12-month period. Having purchased a pair of shoes in one quarter does not preclude a CU from purchasing shoes in some subsequent quarter within the four-quarter period. Likewise, it is not uncommon for a shoe nonpurchaser in one quarter to purchase shoes in a later quarter during a 12-month period.

Let $M = 1$ denote a stayer-purchaser, $M = 2$ a stayer-nonpurchaser, and $M = 3$ a mover. In other words, $M = 1$ denotes a CU who purchases the commodity of interest in each of the four quarters with probability 1; $M = 2$ denotes that a CU has 0 probability of ever purchasing the commodity during the four-quarter window; and $M = 3$ denotes the group consisting CUs who have positive probability of purchasing and not purchasing the commodity in each quarter. Note that M is a *partially* latent variable because it cannot be fully observed. Movers can make purchases of the commodity in all four quarters or no purchases in all four quarters simply by chance. However, by definition, a CU that makes a purchase in 1, 2, or 3 quarters cannot be a stayer.

To reflect this structure in the model, the following constraints are imposed on the purchase status latent variables conditionally on M:

(a) if $M = 1$, then $\Pr(W=1) = \Pr(X=1) = \Pr(Y=1) = \Pr(Z=1) = 1$;

(b) if $M = 2$, then $\Pr(W=1) = \Pr(X=1) = \Pr(Y=1) = \Pr(Z=1) = 0$; and

(c) if $M = 3$, $\Pr(W)$, $\Pr(X)$, $\Pr(Y)$, and $\Pr(Z)$ are unconstrained.

As shown by Langeheine and Van de Pol (1990), the mover-stayer model is identifiable.

Based upon analysis by Biemer (2000), the probability that a respondent falsely reports a purchase is essentially 0. One possible explanation for this is the follow-up questions that are asked about the details of each purchase. If a respondent mistakenly reports a purchase, questions about the particulars of the purchase often reveal the purchase to be in error, in which case the respondent corrects the error. In addition, given the length of the interview, respondents want to avoid the burden of additional questions and, thus, are motivated to only report real purchases. Thus, we may assume

$$\pi_{a=1|w=2} = \pi_{b=1|x=2} = \pi_{c=1|y=2} = \pi_{d=1|z=2} = 0. \qquad (11)$$

In this situation, the latent variables W, X, Y, and Z are only "partially" latent because when they take on the value 0, they are identical to the manifest variable. Partially latent mover-stayer models are discussed in Langeheine and van der Pol (2002). This assumption can be tested in the analysis by freeing the probabilities of a false positive and observing any changes in the fit of the model as well as the effect of removing this constraint on the parameter estimates. However, as we shall see, even when the zero probability constraints are not imposed, the probability of a false purchase report is practically 0 for all commodities in the analysis.

4 Model Selection and Evaluation

Models are estimated separately for each commodity using the ℓEM software. For each item, second-order Markov models were specified with mover-stayer constraints. Grouping variables were chosen from a pool of potential exogenous variables thought to be highly correlated with both purchasing behavior and/or reporting accuracy. For each item, the best model was determined by the set of grouping variables that provided the best fit based on both objective and subjective criteria described subsequently. For the most part, the models easily reached convergence (defined as improvement in maximum log-likelihood of less than 10^{-6}) within 30,000 to 40,000 iterations.[4]

In the course of building these models, the false positive assumption was evaluated. When the constraint was relaxed, most of the false positive estimates produced by the model were zero or very near to zero. They were so near that it was difficult for the models to converge. Thus, the constraint was included in the modeling, and, as a result, the degrees of freedom increased. In addition, the ICE assumption was evaluated by adding direct effects (i.e., interactions between A and B, B and C, or C and D) to the model. Model fit was not significantly improved and no evidence of ICE failure was found.

Among the grouping variables explored in the analysis, the variables providing the best fit to the data were family size, age of householder, household income, and education of householder. In addition, a composite variable formed by combining the use of records with the length of the interview (RECLEN) also improved model fit. This variable together with three other variables were included in the final model.[5] Moving beyond four grouping variables caused convergence problems and improper solutions as a result of data sparseness and overfitting the data. Table 1 provides greater detail about the grouping variables and their coding.

Three methods were used to evaluate the validity of the models. First, the adequacy of the model fit for each commodity was verified using appropriate model diagnostics. Second, as previously stated, several new subjective diagnostic techniques were applied to check the plausibility (face validity) of the estimates produced by the model. Finally, where possible, we compared the model estimates to the external estimates that are considered to be the best estimates available. These comparisons provide some measure of the criterion validity of the results.

Table 2 shows the L^2, Bayesian Information Criterion (BIC), and dissimilarity index for the final models for each commodity, where L^2 is the likelihood ratio chi-square goodness-of-fit statistic. An L^2 having a p-value of 0.05 or higher indicates the hypothesis of model fit cannot be rejected at the 5 percent alpha-level. Generally, the higher the p-value, the better the model fit. BIC = $L^2 - df \ln n$, where df denotes the model degrees of freedom and n is the sample size. It is designed to balance the improvement in L^2 as a result of adding more parameters to the model against the objective of model parsimony. Finally, the dissimilarity index, D, is given by $D = \sum_i |n_i - \hat{m}_i| / (2n)$, where n_i is the observed count in cell i and $\hat{m}_i$ is the expected count under the model. D may be interpreted as the proportion of the data that would have to change cells in order for the model to fit perfectly. For very large data sets, the power of the L^2 test is very high, and models differing trivially from the observed data will be rejected. In that case, D is the preferred measure of fit because it does not depend upon the sample size. A model having a dissimilarity index below 0.05 is considered to fit the data adequately regardless of how small the $L^2 p$-value. Thus, in our analysis, the model selected had the smallest value of BIC among models that fit the data well (i.e., $p > 0.05$ or $D < 0.05$).

Note that for all the commodities in Table 2, D is less than 0.05. Further, except for the three commodities–drugs and medical supplies; trash collection; and televisions, video, and sound equipment–the p-value for the L^2 statistic is at least 0.1. For most items, then, departures from the model are negligible.

Next, several new subjective diagnostic approaches were applied to the results to ascertain face validity (i.e., plausibility of the results). Table 3, column 1 lists all possi-

[4] Starting values were also used to expedite convergence. These starting values were based on the simple measurement model.

[5] Grouping variables, especially those of a demographic nature (nonmethodological), were subject to adjustments by staff of the Office of Prices and Living Conditions at BLS. Adjustments are made if the variable value is missing or fails a consistency check. At this point, values are either directly derived or imputed based on information about the CU itself across the entire panel, or the percentage distributions of all CUs in the sample.

Table 1: Exogenous Variable Definitions

Name	Description	Value	Description	
FAMSIZE	Size of consumer unit (CU)	1	One person CU	
	(Number of missing = 0)*	2	Two person CU	
		3	Three or more person CU	
AGE	Age of respondent	1	Less than 30	
	(Number of missing = 0)*	2	Age 30 to 49	
		3	Age 50 or older	
EDUC	Education of respondent	1	Less than high school	
	(Number of missing = 0)*	2	High school or more	
INCLASS	Income percentile	1	Less than 25th percentile	
	(Number of missing = 0)*	2	25th to 75th percentile	
		3	Greater than 75th percentile	
TYPEREC	Type and frequency of records used during interview (coded by interviewer) (Number of missing = 545)	1	Never or almost never used records	
		2	A single type of record and/or mostly or occasionally used records & missing	
		3	Multiple types of records used almost always or always	
INTLEN	Length of interview (coded by interviewer)	1	Less than 45 minutes	
	(Number of missing = 228)	2	At least 45 minutes, but less than 90 & missing	
		3	At least 90 minutes	
RECLEN	Combined variable of TYPEREC and INTLEN		INTLEN	TYPEREC
		1	1	1
		1	3	1
		1	3	2
		1	3	3
		2	2	1
		3	1	2
		3	2	2
		3	1	3
		3	2	3

* For these variables, missing values were replaced by imputed values using automatic edits and checks based on the CUs' responses on all survey panel waves. This was conducted by the staff of the Office of Prices and Living Conditions at the Bureau of Labor Statistics.

ble reporting patterns, *ABCD,* and columns 2 and 3 provide the observed proportion and number, respectively, of cases with that reporting pattern. The next four columns provide the model estimate of the probability of truly purchasing the commodity at each of the four time periods, given the observed pattern in column 2. Finally, the last column provides the model estimate of the proportion in the population with the true pattern given in column 1. These data were analyzed for all the commodities in Table 2 under the best-fitting model.

Table 3 illustrates this idea for one commodity—shoes. For example, consider a group in the sample reporting the pattern 1121, which corresponds to a purchase of shoes in periods 1, 2, and 4, but not in period 3. In Table 3, locate the pattern 1121 in column 1, third row down from the header. From the second column, note that about 5 percent of respondents (which is the 445 respondents in column 3) reported purchasing shoes in all but the third period. The fourth through seventh columns provide the estimated probabilities that a respondent truly purchased shoes in the first through fourth periods of the study, given they reported the purchase pattern. Recall that the model is constrained so that the probability of a false positive report of a purchase is 0. Therefore, for all but the third period, the probability that a respondent with the pattern 1121 reported a purchase is 100 percent. Even in the third period, more than 80 percent of respondents are estimated as being true purchasers. The eighth column labeled Pr(*WXYZ*) shows the proportion of all respondents that are estimated to have *truly* purchased only in periods 1, 2, and 4 (i.e., 2 percent of the population). Thus, we see that 5 percent of respondents reported pattern 1121, but only 2 percent truly should have reported this pattern according to the model. Note also that the largest observed cell, other than the first and last, is the one in which purchases for shoes are reported in the first quarter but never again. This finding supports our conjecture that after completing the survey once, respondents become more likely to not want to report expenditures in the following interviews.

Table 2: Exogenous Variable Definitions

Commodity	L^2	p-value	BIC	Dissimilarity index
Clothing accessories	329.63	1.00	-5,248.22	0.03
Cable/satellite television	68.27	0.13	-440.46	0.01
Clothing	35.92	0.38	-272.95	0.02
Dental care	164.64	0.30	-1,252.54	0.03
Drugs and medical supplies	72.98	0.00	-235.89	0.02
Electricity	159.10	0.42	-1,258.07	0.01
Eye care	145.58	0.71	-1,271.59	0.02
Furniture	580.16	0.83	-4,997.68	0.05
Gas (household)	68.28	0.13	-440.45	0.01
Kitchen accessories	239.48	0.76	-2,086.13	0.04
Other household items	399.66	1.00	-5,178.19	0.05
Pets and pet supplies	446.63	1.00	-5,131.21	0.04
Shoes	157.90	0.44	-1,259.27	0.04
Sports equipment	391.43	1.00	-5,186.42	0.03
Trash collection	72.00	0.07	-436.73	0.01
Televisions, video, and sound equipment	70.01	0.10	-438.72	0.02
Vehicle service, major	49.16	0.73	-459.57	0.01
Vehicle service, minor	46.27	0.82	-462.46	0.02
Vehicle service, oil changes only	141.98	0.78	-1,275.19	0.03
Vehicle expenses, other	57.84	0.41	-450.89	0.01

BIC = Bayesian Information Criterion

Of course, most of the proportions (except for the first cell and, to a lesser extent, the last), change very little in absolute terms. Instead, it is the proportional change that matters here. As often happens in other fields (such as astronomy or economics), seemingly small shifts can add up to large estimation problems.

Table 4, which is for the same shoes model, shows the proportion of respondents classified as a mover, stayer-purchaser, or stayer-nonpurchaser, given their patterns of reported expenditure. Once again, because we fixed the probabilities of false positives to zero, respondents who reported a purchase in one or more periods cannot be classified as stayer-nonpurchasers by the model. The third row contains those respondents who reported purchasing in all but the third period (1121). About 80 percent of these respondents are classified as stayer-purchasers (i.e., the probability that a CU with a 1121 pattern truly has a purchasing pattern of 1111 is approximately 80 percent). Thus, only 20 percent of respondents with a 1121 pattern are expected to truly have that pattern according to the model. The model classifies those in the last cell (2222), in which no expenditure for shoes was reported in any quarter, in a way that indicates most of these respondents should be movers, reporting a purchase in at least one of the quarters but certainly not in all. This coincides with the fact that many of those in cells in which an expenditure is reported only once also are classified as movers.

Table 4 suggests that the model produces plausible probabilities for most cell patterns. As an example, note that as nonpurchasing increases, so does the proportion identified as movers. At the same time, the model, taking into account both the number of quarters with no purchases and their sequence, identifies many of the so-called movers (based on actual reporting) as stayer/purchasers. Of course, this situation results in underreports in a number of categories, both quarterly and annually.

Table 5 shows the distributions of several key probabilities by the grouping variables, again for the shoes commodity. Three grouping variables are used in the best model for shoes: family size, education, and RECLEN. The probabilities shown in the first section of this table are accuracy rates denoted by $P(A=1|W=1)$. The accuracy rate for a period is the estimated probability that a CU reported purchasing the item in the period, given that it truly purchased the item in the period; for example, in period 1, the accuracy rate is given by the model estimate of $\pi_{A=1|W=1}$. Depending on the subgroup, accuracy rates range from 37 percent to 80 percent. Also included are the percentages of true purchasers (denoted by $P(W=1)$) and reported purchasers (denoted by $P(A=1)$) broken out by the grouping variables, as well as the cell sizes.

The data in Table 5 were produced for each model and were used extensively as a diagnostic metric. Relationships of the accuracy rates by grouping variables were examined for face validity (i.e., whether groups and commodities that are expected to have high reporting accuracy actually do so under the model and vice versa). The relationship of the grouping variables to the accuracy of reporting varies across items, although the use of records and length of interview are both consistently positively related to accuracy. Family size also is positively associated with accuracy in reporting, especially regarding items that are purchased more regularly by larger families (e.g., clothing, dental care, furniture, kitchen accessories, pets and pet supplies). Single-person households tend to have the lowest reporting accuracy which may have more to do with their demographic characteristics than their living arrangements.

Table 3: Probability of Being a True Purchaser by Pattern of Reporting: Shoes

			Entries are Pr(true purchaser | *ABCD*)					
Pattern	Pr(*ABCD*)	n(*ABCD*)	*W*	*X*	*Y*	*Z*	Pr(*WXYZ*)	*n(WXYZ)*
1111	0.09	793	1.00	1.00	1.00	1.00	0.53	4,715
1112	0.05	448	1.00	1.00	1.00	0.85	0.02	168
1121	0.05	445	1.00	1.00	0.85	1.00	0.02	170
1122	0.04	364	1.00	1.00	0.72	0.72	0.02	167
1211	0.05	436	1.00	0.87	1.00	1.00	0.02	141
1212	0.04	389	1.00	0.67	1.00	0.66	0.03	255
1221	0.04	377	1.00	0.69	0.68	1.00	0.02	217
1222	0.07	603	1.00	0.41	0.44	0.43	0.05	424
2111	0.05	434	0.87	1.00	1.00	1.00	0.02	138
2112	0.04	360	0.73	1.00	1.00	0.72	0.02	168
2121	0.04	370	0.71	1.00	0.70	1.00	0.02	205
2122	0.06	557	0.44	1.00	0.44	0.46	0.04	388
2211	0.04	326	0.79	0.79	1.00	1.00	0.01	104
2212	0.06	511	0.52	0.48	1.00	0.45	0.04	314
2221	0.05	480	0.54	0.53	0.48	1.00	0.03	243
2222	0.22	1,923	0.25	0.23	0.21	0.19	0.11	1,002

Table 4: Mover-Stayer Class by Pattern of Reporting: Shoes

	Pr(M|*ABCD*)		
Pattern	Stayer/purchaser	Stayer/nonpurchaser	Mover
1111	0.92	0.00	0.08
1112	0.80	0.00	0.20
1121	0.80	0.00	0.20
1122	0.63	0.00	0.37
1211	0.82	0.00	0.18
1212	0.59	0.00	0.41
1221	0.61	0.00	0.39
1222	0.32	0.00	0.68
2111	0.82	0.00	0.18
2112	0.64	0.00	0.36
2121	0.62	0.00	0.38
2122	0.34	0.00	0.66
2211	0.71	0.00	0.29
2212	0.38	0.00	0.62
2221	0.40	0.00	0.60
2222	0.10	0.16	0.73

The model indicates that respondents with higher education tend to report more accurately. However, those with less than a high school diploma tend to have better accuracy on items that are generally difficult to report (i.e., accessories, eye care, furniture, major vehicle purchases [Veqmaj], miscellaneous vehicle purchases [Veqoth]). Many of these respondents are elderly, and this is consistent with Tucker (1992), who found that elderly respondents were the most conscientious reporters in the Diary Survey component of the CE. It probably is true that their recall is better for these types of items because they have fewer purchases in these categories. Overall, however, the model further indicates that respondents aged 30–49 report purchases more accurately than both older and younger respondents, but younger people appear to report major household items (e.g., furniture or televisions, video, and sound equipment) more accurately.

Respondents with higher income are typically more accurate reporters according to the model. For items that are purchased infrequently by lower income households, both the top and bottom 25 percent of the population classified by income have higher accuracy rates than the middle 50 percent.

4.1 Results for Other Commodities

As a third and final check on the validity of the model estimates, we compared the estimates of reporting accuracy with PCE information. In addition to personal expenditures, the PCE also includes expenditures made on behalf of households by nonprofit organizations and government programs, such as Medicare and Medicaid. Unlike the CE, the PCE excludes person-to-person transactions. The data for the PCE

Table 5: Reporting Accuracy, Cell Size, Percent True and Reported Purchasers by Family Size (*FS*), Education (*E*), and RECLEN (*R*) for Commodity: Shoes

FS = 1

P(*A* = 1 | *W* = 1) (Accuracy)				
	R = 1	*R* = 2	*R* = 3	
E = 1	59.0	54.7	37.3	42.7
E = 2	62.8	58.6	41.2	47.8
	62.2	57.7	40.2	46.6

Cell sizes				
	R = 1	*R* = 2	*R* = 3	Total
E = 1	99	145	463	707
E = 2	289	366	1,088	1,743
	388	511	1,551	2,450

P(*W* = 1) (True Purchaser)				
	R = 1	*R* = 2	*R* = 3	
E = 1	0.6	1.1	4.3	6.0
E = 2	2.9	3.6	12.5	19.0
Total	3.5	4.7	16.8	25.1

P(*A* = 1) (Reported Purchaser)				
	R = 1	*R* = 2	*R* = 3	Total
E = 1	0.6	1.0	2.6	4.2
E = 2	3.0	3.4	8.4	14.8
Total	3.5	4.4	11.0	18.9

FS = 2

P(*A* = 1 | *W* = 1) (Accuracy)				
	R = 1	*R* = 2	*R* = 3	
E = 1	70.2	66.4	49.4	56.9
E = 2	73.5	69.9	53.4	61.4
Total	72.8	69.1	52.5	60.4

Cell sizes				
	R = 1	*R* = 2	*R* = 3	Total
E = 1	108	124	326	558
E = 2	447	432	967	1,846
Total	555	556	1,293	2,404

P(*W* = 1) (True Purchaser)				
	R = 1	*R* = 2	*R* = 3	
E = 1	1.0	1.3	3.4	5.6
E = 2	4.2	4.2	10.8	19.3
Total	5.2	5.5	14.2	24.9

P(*A* = 1) (Reported Purchaser)				
	R = 1	*R* = 2	*R* = 3	Total
E = 1	1.1	1.4	2.7	5.2
E = 2	5.0	4.8	9.4	19.2
Total	6.1	6.2	12.1	24.4

FS = 3

P(*A* = 1 | *W* = 1) (Accuracy)				
	R = 1	*R* = 2	*R* = 3	
E = 1	78.7	75.6	60.5	65.9
E = 2	81.3	78.5	64.4	71.0
Total	80.8	78.0	63.3	69.8

Cell sizes				
	R = 1	*R* = 2	*R* = 3	Total
E = 1	207	113	690	1,010
E = 2	711	525	1,717	2,953
Total	918	638	2,407	3,963

P(*W* = 1) (True Purchaser)				
	R = 1	*R* = 2	*R* = 3	
E = 1	2.5	1.4	8.6	12.5
E = 2	9.6	6.2	21.7	37.5
Total	12.2	7.6	30.3	50.1

P(*A* = 1) (Reported Purchaser)				
	R = 1	*R* = 2	*R* = 3	Total
E = 1	3.2	1.7	8.4	13.4
E = 2	12.7	7.9	22.7	43.3
Total	15.9	9.6	31.1	56.7

are collected from a number of surveys, most significantly the Economic Census, which is conducted by the Census Bureau every 5 years. To produce the PCE, the BEA collects receipts from establishments indicating the value of commodities (services and durables), as well as data to estimate taxes, transportation cost, and trade margins. This value is apportioned to the various sectors to which the commodity is sold: government, exporters, and industry, with the residual allocated to the household sector and the PCE. The ratios of CE estimates of expenditures to the PCE estimates in comparable categories have been used by BLS to provide an independent assessment of the accuracy of the CEs. Other than MLCA approach, it is the only currently available means for checking the accuracy of the CE results. Unfortunately, PCE estimates are limited to just a few commodities. Table 6 contains the ratios for the time period our data cover.

Besides these ratios, Table 6 also shows the model estimates of reporting accuracy, the observed purchase prevalence, and the estimated true purchase prevalence for each commodity. Under the model, $1 - \hat{\pi}^{A|W}_{1|1}$ (i.e., one minus the accuracy rate) is a useful summary statistic for describing the degree of underreporting of purchases. The comparisons of CE to PCE described earlier are presented when the categories in both can be matched. For this particular exercise, the PCE estimates are interpreted as gold-standard estimates.

Apparently, a large degree of variation exists in reporting accuracy across commodities, with the highest accuracy at close to perfect (99.4 percent) and the lowest accuracy at 28.7 percent. The most accurately reported items appear to be those that are purchased regularly, such as electricity (99.4 percent), gas (99.3 percent), cable or satellite television (99.2 percent), and trash collection (94.9 percent) or items such as drugs and medical supplies (94.2 percent). Accuracy also appears to be correlated with the regularity with which a commodity is purchased in the population (0.80 correlation) and to a lesser extent, with the size of the purchase (0.45 correlation) using the MLCA estimate of prevalence as the indicator of purchase frequency.

The highest accuracy rates estimated from the model coincide with the highest CE/PCE ratios in all but one case (pets and pet supplies). The other available ratios, except for furniture, correspond reasonably well, but they can be substantially smaller than 1.0, particularly in the case of clothing accessories. Furniture is one of the five cases in which the model does not fit well. Three out of the other four have no match in PCE, and it is quite possible the same incongruence would be found for those categories.

4.2 Expenditure Estimation

For the purposes of this paper, the estimation of a missing expenditure is confined to the false negative reporters in the sample (i.e., persons who report no expenditure of a particular type but the model predicts otherwise). As previously noted, false positive reporting (i.e., reporting a nonexistent expenditure) is believed to be inconsequential based upon prior research. Furthermore, the problem of inaccuracies in expenditures actually reported, although quite important, is not considered in this paper. That problem is much more difficult and the current methodology is not intended to address it.

The method for estimating the total expenditure missing from a CE report as a result of the failure to report any expenditure for a commodity involves two-steps. First, for each item in Table 2 and using the best fitting MLCA model, we estimate the proportion of false negative reports in group g; for example,$\hat{\pi}_{A=2|W=1,G=g}$for quarter 1. This quantity is used to derive an estimate of the proportion of respondents who reported no expenditure but should have reported at one, denoted by$\pi_{W=1|A=2,G=g}$. It can be easily shown that this estimator is

$$\hat{\pi}_{W=1|A=2,G=g} = \hat{\pi}_{A=2|W=1,G=g} \frac{\hat{\pi}_{W=1|G=g}}{\hat{\pi}_{A=2|G=g}}. \tag{12}$$

LetS_{Ag}denote the set of CUs in group g that responded in the first quarter, and let $\hat{N}_{Ag}$ denote the sum of the base weights (i.e., inverse probabilities of selection in the sample) for persons in S_{Ag}, which is an estimator of the number of persons in the population represented by S_{Ag}. Let $S_{Ag}^{(R)}$denote the subset of S_{Ag}corresponding to persons reporting an expenditure for the commodity of interest. Let $S_{Ag}^{(N)}$ denote persons who did not report an expenditure and note that $S_{Ag} = S_{Ag}^{(R)} \cup S_{Ag}^{(N)}$. Likewise, let $\hat{N}_{Ag}^{(R)}$ and $\hat{N}_{Ag}^{(N)}$ denote the population sizes corresponding to$S_{Ag}^{(R)}$ and $S_{Ag}^{(N)}$, respectively, estimated by summing the weights for all units in $S_{Ag}^{(R)}$ and $S_{Ag}^{(N)}$, respectively, and note that $\hat{N}_{Ag} = \hat{N}_{Ag}^{(R)} + \hat{N}_{Ag}^{(N)}$.

Let $\hat{Y}_{Ag}^{(R)}$denote the total expenditure of CUs in group g estimated by weighting the expenditures of persons in $S_{Ag}^{(R)}$ by their base weights. That is,

$$\hat{Y}_{Ag}^{(R)} = \sum_{i \in S_{Ag}^{(R)}} \omega_{Agi} y_{Agi}, \tag{13}$$

where y_{Agi} is the expenditure for the ith person in group g in the first quarter and $\omega_{Agi}$is that person's corresponding base weight. Then an estimator of the missing expenditure for a person in $S_{Ag}^{(N)}$ is

$$\hat{Y}_{Ag}^{(mis)} = \hat{\pi}_{W=1|A=2,G=g} \hat{N}_{Ag}^{(N)} \frac{\hat{Y}_{Ag}^{(R)}}{\hat{N}_{Ag}^{(R)}}. \tag{14}$$

Tacitly assumed in (14) is that the average expenditure for persons in $S_{Ag}^{(N)}$ who should have reported an expenditure is the same as the average expenditure for persons in $S_{Ag}^{(R)}$. Thus, expenditures within a stratum are considered "missing at random". Finally, an estimate of the total expenditure for the period adjusted for false negatives is

$$\hat{Y}_{Ag}^{(adj)} = \hat{Y}_{Ag}^{(R)} + \hat{Y}_{Ag}^{(mis)}. \tag{15}$$

This estimation process is repeated for all four quarters. The estimate of the total expenditure for all four time periods for group g is $\hat{Y}_{g}^{(adj)} = \hat{Y}_{Ag}^{(adj)} + \hat{Y}_{Bg}^{(adj)} + \hat{Y}_{Cg}^{(adj)} + \hat{Y}_{Dg}^{(adj)}$, and the estimate for all groups is $\hat{Y}^{adj} = \sum_{g} \hat{Y}_{g}^{(adj)}$.

In the example of the shoes commodity, the accuracy rate for respondents living alone, with a high school degree or more education, and who used records but had a relatively short interview is 0.586. The weighted mean reported expenditure for an average quarter for all members of this group who reported at least one expenditure for shoes is \$23.63. After adjusting for nonreporters using (9), the mean becomes \$41.72. Combining it with the other g group or cell means defined by the grouping variables after adjusting in the same manner, the result is a nonreporter adjusted mean of \$61.10 for the entire sample. Table 7 lists, for all commodities, the unadjusted means, the means adjusted for accuracy of reporting, and the relative bias of the average quarter means, treating the adjusted mean as the gold standard. The table indicates that the largest relative biases occur for purchases that are made on an irregular basis, such as vehicle repairs, clothing accessories, furniture, and eye care. In contrast, regular or frequent purchases such as electricity, cable, gas, and

Table 6: Percent with Reported Purchase, True Purchasers, Accuracy Rates, and Consumer Expenditure/Personal Consumption Estimates (CE/PCE) Ratios for Matched Categories

Consumer item	Percent with reported purchase	Percent with true purchase	Accuracy rates	Unadjusted CE/PCE expenditure ratio†
Electricity	76.6	77.1	99.4	1.00
Gas (housing unit)	76.6	77.1	99.3	0.87
Cable/satellite TV	63.0	63.5	99.2	0.93
Trash collection	22.3	23.5	94.9	1.06
Drugs and medical supplies	50.1	53.2	94.2	
Clothing	67.4	75.9	88.8	0.54‡
Vehicle service, oil changes only	39.9	53.4	74.8	
Dental care	22.8	31.6	72.2	
Televisions, video, & sound equipment	35.8	53.6	66.9	0.52
Kitchen accessories	27.7	41.4	66.8	
Shoes	43.7	71.0	61.6	0.74
Pets and pet supplies	14.9	25.0	59.8	0.94
Vehicle expenses, other	10.4	18.5	56.4	
Sports equipment	17.4	32.7	53.1	
Vehicle service, major	15.5	31.8	48.7*	
Vehicle service, minor	26.7	56.4	47.4*	
Eye care	12.2	27.3	44.8*	
Other household items	23.0	55.6	41.4	
Clothing accessories	14.1	40.5	34.7*	0.22‡
Furniture	15.0	52.2	28.7*	0.74

* Poorly fitting models. For details on model fit refer to Table 2.

† Due to large variability in the CE/PCE ratio, an average for the years 1992, 1997, and 2002 is used.

‡ PCE combines clothing and accessories, but the CE/PCE ratios are adjusted based upon additional analysis that provides information on the relative underreporting of each.
Note: In most cases small differences exist in the definition or contents of the consumer item category in the CE/PCE ratio and those used in the analysis.

trash collection tend to have very small (less than 5 percent) relative biases.

Table 7 also contains the CE/PCE ratios, but, in this case, the CE expenditures used are the adjusted ones. The adjustments made appear to work best for frequently purchased items. The estimate for furniture clearly is suspect, but again the model fit was poor. An overadjustment for pets and pet supplies also appears likely. For clothing and clothing accessories, the ratios are the same, because the quality of the adjustments is assumed to be the same.

5 Discussion

It has been long suspected but virtually impossible to quantify that survey estimates of expenditures are biased downward. This paper shows that MLCA may provide useful indicators of the magnitudes of the underreporting problems for most commodities. In addition, MLCA provides insights regarding the correlates of expenditure underreporting that can lead to improved survey design and greater reporting accuracy. This analysis of the CEIS demonstrates the potential of this modeling tool for evaluating the bias in panel survey data.

The analysis found that family size, education, the use of expenditure records, and length of interview are strong determinants of CE reporting accuracy. In general, the use of records increases reporting accuracy. Longer interviews are usually associated with greater accuracy, although whether this is a causal factor is a subject of some debate. Good reporters also tend to be better educated and reside in larger families. This study as well as prior research (for example, Biemer and Tucker 2001) also found that reporting was better for commodities that are purchased regularly rather than infrequently and erratically. Biemer and Tucker also found that accuracy tends to decline as the interview progresses. Of course, all these findings need to be further investigated and verified by field work, but the implications for improving survey methodology are obvious.

This approach also found that the magnitude of the bias in expenditure reports can be substantial for some commodities. For example, one of the largest items for purchase is furniture, yet it is among the most underreported items, with an estimated relative bias of more than 70 percent. However, the underreporting bias for household utilities is usually less than 5 percent. If these results hold true, then more effort should be exerted toward obtaining more accurate data for those commodities exhibiting the largest underreporting biases.

Future research will combine data from more than 3 years of the CEIS in order to estimate more complex models with more grouping variables, including underreporting among those having an expenditure. We hypothesize that respondents who report a purchase are likely to underestimate the amount of that purchase, especially over a 3-month time period. We plan to develop a model at the micro level to estimate the underreporting of the expenditure by an individual respondent who actually reports an expenditure. This measure will be a latent variable that combines several indicators of the respondent's level of effort in the survey. It also could

Table 7: Mean First Quarter and "Average" Quarter Expenditure for All Survey Respondents[1] and the CE/PCE Ratios with Adjusted Expenditures for Matched Categories

	First Quarter		Average Quarter			
Consumer item	Adjusted	Unadjusted	Adjusted	Unadjusted	Relative Bias (%)	Adjusted CE/PCE Ratio †
Clothing accessories*	15.60	5.83	14.89	5.36	-64.0	0.62
Cable/satellite TV	66.04	65.49	69.38	68.84	-0.8	0.94
Clothing	213.48	189.56	198.52	174.85	-11.9	0.62
Dental care	84.54	60.82	81.43	58.37	-28.3	
Drugs and medical supplies	72.40	68.13	74.96	70.76	-5.6	
Electricity	195.71	194.43	198.26	196.96	-0.7	1.01
Eye care*	42.52	19.34	42.40	18.74	-55.8	
Furniture*	275.24	80.68	253.86	71.44	-71.9	2.63
Gas (household)	67.41	66.81	66.12	65.53	-0.9	0.88
Kitchen accessories	56.33	37.02	49.55	32.07	-35.3	
Other household items	85.60	38.07	82.80	35.23	-57.5	
Pets and pet supplies	27.96	17.25	30.74	19.19	-37.6	1.51
Shoes	61.37	38.17	61.10	37.40	-38.8	1.21
Sports equipment	57.86	32.24	54.98	29.30	-46.7	
Trash collection	11.37	10.84	11.72	11.20	-4.4	1.11
Televisions, video, & sound equipment	154.13	103.88	143.14	92.17	-35.6	0.81
Vehicle service, major*	110.62	58.42	104.26	53.42	-48.8	
Vehicle service, minor*	120.64	57.66	111.86	51.86	-53.6	
Vehicle service, oil changes only	20.58	15.58	20.33	15.19	-25.3	
Vehicle expenses, other	22.74	12.37	21.48	11.56	-46.2	

[1] n = 8,817
* Poorly fitting models. For details on model fit refer to Table 2.
† Due to large variability in the CE/PCE ratio, an average for the years 1992, 1997, 2002 is used.
Note: In most cases small differences exist in the definition or contents of the consumer item category in the CE/PCE ratio and those used in the analysis.

offer an opportunity to further investigate individual cases of false positives.

We will begin with one quarter of data and later develop a measure based on all four quarters of data. In using all four quarters, we will not exclude respondents who fail to complete one or more interviews. This information will be valuable for evaluating the expenditure reports in the interviews they do complete.

Regression models will be developed to determine what respondent characteristics are most strongly associated with the underreporting. A method of estimation for each commodity also using regression models, with both the latent variable and demographic variables, will adjust the expenditure level for each underreporter. Later work will combine all results to produce a more complete explanation of the measurement error in the reporting of purchases on the CEIS, by both underreporters and nonreporters, and adjustments for that error.

References

Biemer, P. P. (2000, June). *An Application of Markov Latent Class Analysis for Evaluating the Screening Questions in the CE Interview Survey.* Technical report submitted to the Bureau of Labor Statistics, U.S. Department of Labor.

Biemer, P. P. (2011). *Latent Class Analysis of Survey Error*. Hoboken, NJ: John Wiley & Sons.

Biemer, P. P., & Tucker, C. (2001). *Estimation and Correction for Purchase Screening Errors in Expenditure Data: A Markov Latent Class Modeling Approach.* Proceedings of the International Statistics Institute, Seoul, Korea.

Blumen, I. M., Kogan, M., & McCarthy, P. J. (1955). *The Industrial Mobility of Labor as a Probability Process*. Ithaca, NY: Cornell University Press.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, *39*, 1-38.

Goodman, L. A. (1961). Statistical Methods for the Mover-Stayer Model. *Journal of the American Statistical Association*, *56*, 841-868.

Goodman, L. A. (1974). Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, *61*, 215-231.

Langeheine, R., & van de Pol, F. (1990). A Unifying Framework for Markov Modeling in Discrete Space and Discrete Time. *Sociological Methods and Research*, *18*(3), 416-441.

Langeheine, R., & Van der Pol, F. (2002). Latent Markov Chains. In J. Hagenaars & A. McCutcheon (Eds.), *Applied Latent Class Analysis* (p. 305-342). Cambridge, UK: Cambridge University Press.

Lazarsfeld, P. F. (1950). The Logical and Mathematical Foundation of Latent Structure Analysis. In S. Stauffer, E. A. Suchman, P. F. Lazarsfeld, S. A. Starr, and J. Clausen (Ed.), *Studies on Social Psychology in World War II, Vol. 4, Measurement and Prediction.* Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis*. Boston, MA: Houghton-Mifflin.

Tucker, C. (1992). The Estimation of Instrument Effects on Data Quality in the Consumer Expenditure Diary Survey. *Journal of*

Official Statistics, *8*, 41-61.

Tucker, C., Biemer, P., & Vermunt, J. (2002). *Estimation Error in Reports of Consumer Expenditures.* Proceedings of the American Statistical Association, Survey Research Methods Section. New York, NY: American Statistical Association.

Van de Pol, F., & de Leeuw, J. (1986). A Latent Markov Model to Correct for Measurement Error. *Sociological Methods & Research*, *15*(1-2), 118-141.

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

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