Methods to Account for Uncertainty in Latent Class Assignments When Using Latent Classes as Predictors in Regression Models, with Application to Acculturation Strategy Measures

Michael R. Elliott,[a,b] *Zhangchen Zhao,*[a] *Bhramar Mukherjee,*[a] *Alka Kanaya,*[c] *and Belinda L. Needham*[d]

Abstract : Latent class models have become a popular means of summarizing survey questionnaires and other large sets of categorical variables. Often these classes are of primary interest to better understand complex patterns in data. Increasingly, these latent classes are reified into predictors of other outcomes of interests, treating the most likely class as the true class to which an individual belongs even though there is uncertainty in class membership. This uncertainty can be viewed as a form of measurement error in predictors, leading to bias in the estimates of the regression parameters associated with the latent classes. Despite this fact, there is very limited literature treating latent class predictors as measurement error models. Most applications ignore this issue and fit a two-stage model that treats the modal class prediction as truth. Here, we develop two approaches—one likelihood-based, the other Bayesian—to implement a joint model for latent class analysis and outcome prediction. We apply these methods to an analysis of how acculturation behaviors predict depression in South Asian immigrants to the United States. A simulation study gives guidance for when a two-stage model can be safely implemented and when the joint model may be required.

Keywords: Bayesian models; Depression; Latent class analysis; Measurement error

(*Epidemiology* 2020;31: 194–204)

Latent class analysis,[1,2] which is a subset of structural equation modeling, is a model-based, person-centered technique for identifying unobservable subgroups within a population.[3] Latent class analysis is well suited for the analysis of multidimensional theoretical constructs that cannot be measured directly, such as acculturation strategies, socioeconomic status, social networks, and healthy lifestyles, for example. Structural equation modeling has a long history in the social and behavioral sciences, but health researchers have increasingly recognized its potential to enhance epidemiologic research[4]; for a more critical assessment of structural equation modeling in epidemiology, see.[5] Noted strengths of structural equation modeling include the use of multiple observed variables to estimate unobserved, or latent, variables; explicit modeling of measurement error; the ability to simultaneously model complex patterns of relationships; and a confirmatory approach to test how well empirical data fit an underlying theoretical model.[6] Latent class analysis, in particular, is useful when there are differences in the underlying distributions among variables that contribute to heterogeneity (e.g., in risk or response to a treatment) within a study population.[7]

Latent class analysis defines classes by the categorical probabilities of observed (manifest) variables and then assumes that, conditional on an unknown (latent) class membership for each subject, the manifest variables are independent of each other; averaging across these classes then induces the correlations observed among the manifest variables. Thus, in a survey of South Asian immigrants, we might find a “separation” and an “acculturation” class. In the separation class, the probability of both native language use and ethnic spice use are high, whereas in the acculturation class, the probability is lower. Marginally, then, language use and cooking style is correlated. However, *conditional* on being in the separation class, the probability of using native language is high, but, given native language is used, it tells us nothing about use of ethnic spices. Similar reasoning follows for the integration class. Latent class analysis has been increasingly used as a statistical tool in epidemiology in the 21st century: to cite just a few examples from this journal, Jokinen and Scott[8] used latent class analysis to combine information from seven tests of pneumococcus presence to estimate the contribution of *Streptococcus pneumoniae* to the burden of community-acquired pneumonia; Hui et al.[9] used latent class analysis to classify

Submitted September 25, 2018; accepted November 6, 2019.

From the [a]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI; [b]Survey Methodology Program, Institute for Social Research, University of Michigan, Ann Arbor, MI; [c]Department of Medicine, University of California - San Francisco, San Francisco, CA; and [d]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI.

Supported by NIH Grant No. 1R01 HL093009.
The authors report no conflicts of interest.
SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Michael R. Elliott, Department of Biostatistics, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109. E-mail: mrelliot@umich.edu.

ISSN: 1044-3983/20/3102-0194
DOI: 10.1097/EDE.0000000000001139

growth patterns for Hong Kong infants; Lasry et al.[10] used latent class analysis to combine data from five administrative health data sources to estimate the true prevalence of traumatic brain injury and relate this back to each administrative data sources to estimate its sensitivity and specificity.

Simple application of Bayes theorem estimates a posterior probability of latent class membership for each subject. Based on these posterior latent class membership probabilities, latent class memberships can be assigned to each subject and used as predictors of distal outcomes. For example, Harlow et al.[11] used latent class analysis to summarize over 50 symptoms reported by midlife women, ranging from hot flashes and sexual dysfunction to measures of pain and mental health. They found six classes of midlife symptoms in women, ranging from highly symptomatic to nonsymptomatic, with special classes dominated by fatigue and psychological symptoms or by physical health symptoms. These classes were, in turn, used to predict self-reported health status, finding that women from high symptom classes were more likely to report poor health, although the fatigue and psychological symptom class was not so likely to do so. Similarly, Parada et al.[12] used latent class analysis to combine information about behavior and diet into three levels of lifestyle used to estimate mortality risk in breast cancer patients using a Cox proportional hazards model.

It is well known that there are uncertainties in class assignment. When latent class assignments are used as predictors in regression analysis, this uncertainty can be recast as measurement error in the predictor. Predictors with normally distributed measurement error is a long-known issue in regression analysis.[13] Independence between the measure error and the predictor yields regression parameter estimates that are biased toward 0 (bias toward the null), with accompanying variance estimates that are too small compared with the repeated sampling variability of the regression parameters that use the true values. In the categorical setting, more recent work has shown that measurement error is perforce correlated with the true predictors, although the issue of bias (which can be toward the null or even reverse signs) remains as a function of the probability of the category and the probability of the value being misspecified.[14] The uncertainty in the latent class assignment—which can be trivial or substantial, depending on the application, but exists by definition—can thus be expected to impact the results of a regression thus uses latent classes as predictors, and their extensions in generalized growth mixture models.[15] This is in contrast to predictors of the latent classes themselves—for example, body mass index, race/ethnicity, smoking status, education, and income in[11]—which are observed in joint estimation of the latent class analysis models. (Although these models have their own complexities—for example, having the construction of the classes differ depending on the choice of predictors,[16,17] that is not the focus of this manuscript.)

Although regression using variables with measurement error is also a well-studied problem,[18] such methods can be costly in terms of time and statistical expertise to implement, and their use in the latent class analysis setting has been more limited. The main methods that have been proposed include (1) assigning latent classes based on modal posterior probability of membership and treating them as fixed and known,[19] (2) making assignments based on posterior modal values, but then using a weighted likelihood method with weights given by an entropy measure of the correct probability of classification,[20] and (3) a multiple imputation approach based on repeated imputations of the latent class based on the vector of their posterior probabilities of class assignment.[21] Here, we proposed to use a full joint model that incorporates the uncertainty in the latent class analysis into the estimation of the outcome. Because of the complexities in model fitting, we develop both an expectation-maximization (EM) algorithm[22] and a fully Bayesian approach.[23] Such approaches are principled in that they fully account for the measurement error in latent class predictors in the regression model, and they have become rather straightforward to implement with the development of general use Bayesian software.[24]

Here, we consider the use of latent class analysis to summarize 12 indicators of acculturation among a sample of South Asian adults in the San Francisco Bay and the greater Chicago area enrolled as part of the Mediators of Atherosclerosis in South Asians Living in America (MASALA) Study.[25] Needham et al.[26] found three latent classes of acculturation termed separation, assimilation, and integration; treating these classes as predictors of depressive symptoms, Needham et al.[27] found no statistically significant differences in levels of depression between these classes after adjusting for age, sex, religious beliefs, education, occupation, income, place of birth, percent of life lived in the United States, English language proficiency, and antidepressant use. (Further adjustment for discrimination and social support showed a marginally higher level of depression in the separation class.) In addition to assigning each subject to the latent class with the highest posterior probability, treating these classes as observed, this analysis used a multiple imputation approach, generating the latent classes for the regression based on their posterior distribution 1000 times and computing the regression estimates and associated standard errors using Rubin’s combining rules.[28] Here, we extend this approach to consider the joint estimate of latent class and regression models for both continuous and binary measures of depression, using both expectation–maximization and Bayesian methods, and compare these results with those obtained from a standard two-step method. We also conduct a simulation study to determine the degree to which this approach reduced bias and mean square error (MSE) and improves nominal interval coverage over a standard two-step approach. We conclude with a discussion of the potential value of this alternative approach to existing methods, and topics for future research.

METHODS

MASALA Study

The MASALA study[25] recruited 906 individuals of South Asian ancestry into a longitudinal prospective cohort study on the prevalence, correlates, and outcomes associated with subclinical cardiovascular disease. Participants were restricted to be 40–84 years of age and to be free of physician-diagnosed cardiovascular disease. Subjects were recruited from the counties surrounding two research clinics, one in San Francisco, the other in Chicago, from October 2010 to March 2013. A wide variety of health measures were obtained by questionnaire and by clinical and radiological examination. The institutional review boards at the University of California, San Francisco, and Northwestern University approved the study protocol, and all study participants provided written informed consent. As in the study by Needham et al.,[27] we excluded 19 subjects born in the United States and 31 subjects with one of more missing variables from the analysis, leaving n = 856 subjects.

Measures

Depression was measured using the Center for Epidemiologic Studies Depression (CES-D) Scale.[29] Respondents reported how often they experienced 20 symptoms within the past week, such as poor appetite, trouble concentrating, and talking less than usual, leading to a score of 0 (no depressive symptoms) to 60 (maximum score). We considered both the continuous measure and a binary outcome of CES-D $\geq$ 16.

As in the study by Needham et al.,[26] 12 manifest variables were used to estimate the three latent classes of acculturation. These included seven measures of South Asian tradition, with reported categories ranging from “absolutely” to “not at all”: (1) performing religious ceremonies or rituals; (2) serving South Asian sweets for ceremonies or rituals; (3) fasting on specific occasions; (4) living in a joint family; (5) having an arranged marriage; (6) having a staple diet of chapatis, rice, dal, vegetables, and yogurt; and (7) using spices for healing and health. The remaining variables included reports on: (8) how often they fast (from “two to three times per week” to “almost never or never”); (9) what foods they normally eat at home (“only South Asian food,” “mostly South Asian food,” “South Asian and other food about equally,” “mostly other food,” “only other food,” “never eat at home”); (10) what foods they normally eat in restaurants (“only South Asian food,” “mostly South Asian food,” “South Asian and other food about equally,” “mostly other food,” “only other food,” “never eat in restaurants”); (11) how often their family shops at South Asian grocery stores or markets (from “two or three times per week” to “almost never or never”); and (12) which country or culture most of their friends belong to (from “only South Asian” to “only other ethnic groups”).

Finally, factors including age, sex, religion, education, income, occupational status, birth nation, proportion of life lived in the United States, English proficiency, and use of antidepressants were used in the regression models as possible confounders between acculturation strategies and symptoms of depression.

Statistical Analyses

The latent class analysis models assumes that the $i=1,\ldots,n$ subjects belong to one of $j=1,\ldots,J$ classes and that each of the $l=1,\ldots,L$ manifest variables that define the latent classes is a multinomial variable Z_{il} with cell probability π_{klj}, $\sum_{k=1}^{K_l}\pi_{klj}=1$ for all $j,l,$ where K_l is the number of cells for the l th manifest variable. Let C_i define the latent class membership of the i th subject, which is in turn multinomially distributed with marginal probability θ_j, $\sum_{j=1}^{J}\theta_j=1$. This yields the following likelihood $L=\prod_i L_i$, where

$$L_i = L_i(\theta_1,\ldots,\theta_J,\pi_{111},\ldots,\pi_{K_L LJ}) = \sum_{j=1}^{J} P(C_i=j)\prod_l P(Z_{il}\mid C_i=j)$$
$$=\sum_{j=1}^{J}\theta_j\prod_l\prod_k \pi_{klj}^{Z_{ikl}}$$

where Z_{ikl} is an indicator equal to 1 if $Z_{il}=k$ and 0 otherwise.

This model can be fit using an expectation-maximization algorithm, as in the poLCA package in R3.4.1.[30] The posterior probability of class membership can then be computed using Bayes theorem:

$$P(C_i=j\mid Z_{i1},\ldots,Z_{iL}) = \frac{P(Z_{i1},\ldots,Z_{iL}\mid C_i=j)P(C_i=j)}{\sum_{j=1}^{J}P(Z_{i1},\ldots,Z_{iL}\mid C_i=j)P(C_i=j)}$$
$$=\frac{\theta_j\prod_l\prod_k\pi_{klj}^{Z_{ikl}}}{\sum_{j=1}^{J}\theta_j\prod_l\prod_k\pi_{klj}^{Z_{ikl}}}$$

with maximum likelihood estimates $\{\hat{\theta}_j\}$ and $\{\hat{\pi}_{kjl}\}$ replacing the true values to obtain an estimated posterior probability of class membership $\hat{P}(C_i=j\mid Z_{i1},\ldots,Z_{iL})$.

Two-stage method: in the two-stage method, the maximum posterior probability is used to obtain a estimate of the true latent class: $\hat{C}_i=\max_j\left(\hat{P}(C_i=j\mid Z_{i1},\ldots,Z_{iL})\right)$. $\hat{C}_i$ is then used as a multinomial predictor in either a linear model (for continuous CES-D) or a logistical model (for binary measure of depression).

Joint method using expectation–maximization algorithm: in the joint method, the full likelihood combines both the latent class model and the regression model through the latent class indicator C. For the linear regression model, let Y_i be the outcome measure and $\{X_{im}\}$, $m=1,..M$ be the M covariates that are to be adjusted for in the outcome model. We make the following assumptions:

1. The covariates X are predictive only of the outcome Y (i.e., $C\perp X\mid Z$).

2. Conditional on latent class C, manifest variables Z are independent of the outcome—that is, all of the relevant information that Z contains about Y is summarized in C.
3. The outcome Y is normally distributed with a mean equal to a linear combination of the latent classes C and the covariates X: $Y_i \mid C_i = j, X_{i1},..., X_{iM} \sim N(\alpha_j + \sum_m \beta_m X_{im}, \sigma^2)$.

The complete data likelihood is then given by $L = \prod_i L_i$, where

$$L_i = L_i(\theta_1,...,\theta_J, \pi_{111},..., \pi_{K_L LJ}, \beta_1,...,\beta_M, \alpha_1,..,\alpha_J) =$$

$$P(Y_i, Z_{i1},...,Z_{iL}, C_i \mid X_{i1},...,X_{iM}) =$$

$$P(C_i \mid X_{i1},...,X_{iM}) P(Z_{i1},...,Z_{iL} \mid C_i, X_{i1},...,X_{iM})\ X$$

$$P(Y_i \mid Z_{i1},...,Z_{iL}, C_i,\ X_{i1},...,X_{iM}) =$$

$$P(C_i) P(Z_{i1},...,Z_{iL} \mid C_i) P(Y_i \mid C_i, X_{i1},...,X_{iM})$$

$$= \prod_j \left\{ \theta_j \prod_l \prod_k \pi_{klj}^{Z_{ikl}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \begin{pmatrix} y_i - \beta_0 - \alpha_j \\ -\sum_m \beta_m X_{im} \end{pmatrix}^2 \right] \right\}^{I(C_i = j)}$$

A logistic model can be obtained by replacing $P(Y_i \mid C_i, X_{i1},...,X_{iM})$ with

$$\frac{\exp(\beta_0 + \alpha_j + \sum_m \beta_m X_{im})^{Y_i}}{1 + \exp(\beta_0 + \alpha_j + \sum_m \beta_m X_{im})}.$$

This method can be implemented using an EM algorithm,[22] treating the indicators of latent class membership as missing; details are provided in eAppendix; http://links.lww.com/EDE/B623. Inference (95% confidence interval [CI], *p* values) can be obtained using a bootstrap procedure.[31]

Fully Bayesian joint method: Bayesian latent class analysis models have been developed previously,[32,33] but extensions to jointly model the classes and the outcome model conditional on the classes are less developed. Using the same joint modeling assumptions (covariates are predictive only of the outcome; conditional on latent class, manifest variables Z are independent of the outcome), we can define the posterior distribution as

$$P(\theta_1,...,\theta_J, \pi_{111},...,\pi_{K_L LJ}, \beta_1,...,\beta_M, \alpha_1,..,\alpha_J \mid Y, X, Z) \propto$$

$$\prod_i L_i(\theta_1,...,\theta_J, \pi_{111},...,\pi_{K_L LJ}, \beta_1,...,\beta_M, \alpha_1,..,\alpha_J) \prod_j$$

$$P(\theta_j) P(\alpha_j) \prod_k \prod_l \prod_j P(\pi_{klj}) \prod_m P(\beta_m)$$

We use standard conjugate-type priors:

$$(P(\theta_1),...,P(\theta_J))^T \sim \mathrm{DIR}(\lambda_1^\theta,..,\lambda_J^\theta)$$

$$P(\alpha_j) \overset{iid}{\sim} N(0, \sigma_\alpha^2)$$

$$(P(\pi_{1lj}),...,P(\pi_{K_l lj}))^T \sim \mathrm{DIR}(\lambda_{1lj}^\pi,..,\lambda_{K_l lj}^\pi)$$

$$P(\beta_m) \overset{iid}{\sim} N(0, \sigma_\beta^2)$$

This method can be implemented using a combination of Gibbs sampling and Metropolis steps in a Markov Chain Monte Carlo (MCMC) algorithm[23]; here, use the adaptive importance sample algorithm in WINBUGS V1.4.[24] Inference is obtained from the empirical distributions of the draws from the MCMC algorithm.

RESULTS

We fit the two-stage, EM, and Bayesian joint method to compare the results obtained under these approaches, for both the linear and the logistic regression models (continuous vs. binary outcome for depression). To implement the Bayesian joint method, we used the following weakly informative priors, setting $\lambda_1^\theta = \cdots = \lambda_j^\theta = 1/J$, $\lambda_{1lj}^\pi = \cdots \lambda_{K_l lj}^\pi = 1/K_l$ for all j, and $\sigma_\beta^2 = \sigma_\alpha^2 = 10$. We obtained 10,000 draws from the posterior distribution after 5000 burn-in draws, with a thinning rate of 10 (i.e., keeping every 10th draw). Trace plots of the parameter draws are given in eFigures 1–11; http://links.lww.com/EDE/B623; all show good convergence in distribution.

A three-class model was fit, yielding acculturation strategy classes that corresponded to an integration class, an assimilation class, and a separation class. Using only the manifest variable data information (i.e., the two-stage approach), 54% (95% CI: 49%, 59%) of the sample belonged to the integration class, 24% (95% CI: 20%, 28%) to the assimilation class, and 23% (95% CI: 19%, 27%) to the separation class. The Bayesian joint method (both linear and logistic) yielded nearly identical distributions, whereas the expectation–maximization joint methods slightly disfavored the integration class (52%) and favored the assimilation (25%) and separation classes (23%). The three classes were quite distinct, with the assimilation class tending to rank low on all measures, the separation class high, and the integration class intermediate between the two. Dietary factors tended to be somewhat less discriminatory, whereas support for arranged marriages perhaps distinguished the most among the three groups. There was little difference between any of the methods in terms of the formation of the three classes, suggesting that the underlying meaning of the classes is stable irrespective of method. eFigures 12–14; http://links.lww.com/EDE/B623; show the exact distribution of the manifest variable in each of the three latent classes under the two-stage method, along with the equivalent distributions when a joint method is fit, using either the continuous or binary outcome measure of depression, and either the expectation–maximization or the fully Bayesian approach. We included the results when the latent class model alone was fit as well for comparison.

Linear regression: Table 1 shows the result of regressing the acculturation strategy classes on the CES-D measure of depression, adjusted for age (a cubic spline model with a knot at age 60 years), sex, religion, education, income, occupational status, birth nation, proportion of life lived in the United States, and English proficiency. There was little

TABLE 1. Changes in CES-D Associated with Acculturation Strategy Classes

	Two-step Model		EM Joint Model		Bayesian Joint Model	
	Beta	95% CI	Beta	95% CI	Beta	95% CI
Assimilation (vs. integration)	−0.11	(−0.26, 0.04)	−0.09	(−0.26, 0.07)	−0.12	(−0.29, 0.05)
Separation (vs. integration)	−0.01	(−0.15, 0.14)	−0.01	(−0.17, 0.16)	−0.03	(−0.20, 0.14)
Age (per decade)	−0.06	(−0.24, 0.12)	−0.06	(−0.24, 0.12)	−0.08	(−0.21, 0.04)
Age square (per decade)	−0.03	(−0.15, 0.09)	−0.03	(−0.15, 0.09)	−0.01	(−0.09, 0.06)
Age cubic (per decade)	−0.01	(−0.16, 0.13)	−0.01	(−0.16, 0.14)	0.00	(−0.06, 0.06)
Splined age cubic (per decade)	0.06	(−0.25, 0.37)	0.06	(−0.25, 0.37)	0.06	(−0.54, 0.69)
Female (vs. male)	0.06	(−0.07, 0.18)	0.06	(−0.07, 0.18)	0.05	(−0.06, 0.18)
Other religion (vs. Hinduism/Jainism)	0.16	(0.00, 0.31)	0.16	(−0.00, 0.31)	0.16	(0.00, 0.32)
No religion (vs. Hinduism/Jainism)	−0.16	(−0.42, 0.10)	−0.17	(−0.43, 0.09)	−0.16	(−0.42, 0.09)
Less than BA (vs. more than BA)	0.30	(0.08, 0.51)	0.30	(0.08, 0.51)	0.30	(0.08, 0.51)
BA (vs. more than BA)	0.26	(0.12, 0.39)	0.26	(0.12, 0.39)	0.26	(0.13, 0.40)
Unemployed (vs. employed)	0.17	(−0.01, 0.34)	0.17	(−0.01, 0.34)	0.17	(−0.00, 0.34)
Retired (vs. employed)	0.26	(0 .05, 0.47)	0.26	(0.05, 0.46)	0.25	(0.05, 0.46)
Household income (per $1K)	−0.10	(−0.25, 0.06)	−0.10	(−0.26, 0.05)	−0.10	(−0.25, 0.06)
Born other South Asia (vs. India)	0.02	(−0.15, 0.19)	0.02	(−0.15, 0.19)	0.02	(−0.14, 0.19)
% of life lived in the United States (10%)	0.01	(−0.02, 0.05)	0.01	(−0.02, 0.05)	0.01	(−0.02, 0.05)
Speak English fairly well Or poorly/not at all (vs. well or very well)	0.14	(−0.06, 0.35)	0.15	(−0.06, 0.35)	0.16	(−0.05, 0.35)

BA indicates bachelor's degree or college; CES-D, Center for Epidemiologic Studies Depression; CI, confidence interval.

difference between the effects of the acculturation strategy classes in the various methods, although there is a hint that the effects are somewhat stronger and more variable in the joint methods than the two-stage method, as anticipated. The effects of acculturation strategy were not significant for any of the methods.

Logistic regression: Table 2 shows the result of regressing the acculturation strategy classes on the binary measure of depression, adjusted for the same variables as in the linear model. (Because of a less complex relationship between the binary outcome of depression and age than the continuous outcome of CES-D and age, the cubic spline for age was replaced with a linear term.) Here, assimilation appeared to be associated with a considerably reduced risk of depression compared with integration in all of the methods. The effect was somewhat larger and more variable in the expectation–maximization method than in the two-step method, again consistent with our expected findings; for the Bayes method, the effect was similar and somewhat more variable. There was not strong evidence that those pursuing a separation strategy were more or less likely to be depressed compared with those pursing an integration strategy under any of the methods, although the point estimates associated with separation with the joint expectation–maximization and Bayesian methods were further from the null than the two-step method, as anticipated.

SIMULATION STUDY

To better understand how our methods work in a controlled setting, we undertook a small simulation study that paralleled in a somewhat simplified manner the MASALA application (sometimes terms a "plasmode" simulation[34]). First, we generated two X covariates independently for $i = 1,...,n$, where $X_{i1} \sim \text{UNI}(2,8)$ (roughly corresponding to age in decades) and $X_{i2} \sim \text{BIN}(1,0.5)$ (roughly corresponding to sex). We then generated our latent class variable using a probit model for C based on latent, normally distributed C^*: $C_i^* \sim N(0.25X_{i1} - 0.7X_{i2},1)$, where $C^* < 0.875 \Rightarrow (C = 1)$, $0.875 \geq C^* < 1.648 \Rightarrow (C = 2)$, and $C^* > 1.648 \Rightarrow (C = 3)$. (Note that this provides marginal distributions corresponding to $C \sim \text{MULTI}(0.5,0.25,0.250)$, the approximate distributions of the three latent classes in the MASALA example.) The manifest variables $Z_{i1},...,Z_{i10}$ are then generated from $Z_{il} \sim \text{MULTI}(1,\pi_{lj})$, depending on latent class j, where $\pi_{lj} = (\pi_{1lj},...,pi_{5lj})^T$ is in turn drawn from a Dirichlet distribution for each simulation with parameters λ_j, where we consider a high separation model

Highseparation

$$\lambda_1 = (5,4,3,2,1)^T$$
$$\lambda_2 = (1,3,5,3,1)^T$$
$$\lambda_3 = (1,2,3,4,5)^T$$

and a low separation model

Lowseparation

$$\lambda_1 = (3,2.5,2,1.5,1)^T$$
$$\lambda_2 = (1,2,3,2,1)^T$$
$$\lambda_3 = (1,1.5,2,2.5,3)^T$$

TABLE 2. Changes in Log Odds of Depression Associated with Acculturation Strategy Classes

	Two-step Model		EM Joint Model		Bayesian Joint Model	
	Beta	95% CI	Beta	95% CI	Beta	95% CI
Assimilation (vs. integration)	−0.86	(−1.6, −0.24)	−1.0	(−1.8, −0.31)	−0.82	(−1.6, −0.16)
Separation (vs. integration)	−0.16	(−0.66, 0.32)	−0.23	(−0.77, 0.28)	−0.25	(−0.81, 0.30)
Age (per decade)	−0.07	(−0.34, 0.19)	−0.07	(−0.34, 0.19)	−0.07	(−0.32, 0.18)
Female (vs. male)	0.34	(−0.11, 0.78)	0.35	(−0.09, 0.80)	0.34	(−0.10, 0.82)
Other religion (vs. Hinduism/Jainism)	0.80	(0.33, 1.26)	0.81	(0.34, 1.3)	0.82	(0.31, 1.3)
No religion (vs. Hinduism/Jainism)	0.58	(−0.50, 1.50)	0.64	(−0.45, 1.6)	0.40	(−0.72, 1.4)
Less than BA (vs. more than BA)	1.0	(0.39, 1.7)	1.1	(0.42, 1.7)	1.08	(0.43, 1.7)
BA (vs. more than BA)	0.60	(0.13, 1.07)	0.61	(0.14, 1.08)	0.61	(0.15, 1.1)
Unemployed (vs. employed)	0.18	(−0.40, 0.75)	0.19	(−0.40, 0.76)	0.20	(−0.39, 0.78)
Retired (vs. employed)	0.56	(−0.11, 1.23)	0.56	(−0.12, 1.22)	0.54	(−0.10, 1.2)
Household income (per $1K)	−0.07	(−0.70, 0.50)	−0.06	(−0.69, 0.51)	−0.07	(−0.51, 0.36)
Born other South Asia (vs. India)	0.38	(−0.16, 0.89)	0.39	(−0.14, 0.90)	0.38	(−0.16, 0.87)
% of life lived in the United States (per 10%)	0.02	(−0.10, 0.15)	0.03	(−0.10, 0.15)	0.02	(−0.10, 0.14)
Speak English fairly well Or poorly/not at all (vs. well or very well)	0.32	(−0.29, 0.90)	0.33	(−0.27, 0.92)	0.34	(−0.24, 0.84)

BA indicates bachelor's degree or college; CI, confidence interval.

Finally, the outcome variable Y is generated by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \alpha_1 I(C_i = 2) + \alpha_2 I(C_i = 3) + \varepsilon_i, \varepsilon_i \sim N(0,1)$$

for the continuous outcome and

$$P(Y_i = 1) = \frac{\exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \alpha_1 I(C_i = 2) + \alpha_2 I(C_i = 3)\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \alpha_1 I(C_i = 2) + \alpha_2 I(C_i = 3)\right)}$$

for the binary outcome. In the continuous model, $\beta_0 = 10, \beta_1 = -0.5, \beta_2 = 2$. In the binary model, $\beta_0 = -0.6 \beta_1 = -0.2, \beta_2 = 1$ so that the marginal probability for Y approximates 0.25. In both models, $(\alpha_1, \alpha_2) = (-0.5, 0.5)$. We consider samples of size $\text{n} = 100, 200$, and 500 to assess the effects of different degrees for information about the latent classes and thus potentially different effects of measurement error.

For each of the $2 \times 2 \times 3 = 12$ scenarios, 200 simulations were run, and bias, MSE, and nominal 95% coverage probability were obtained for each of the fixed β parameters and the α parameters associated with the latent classes. A two-stage, single imputation, and joint expectation–maximization and Bayesian modeling approach are considered. For the joint Bayesian method, the same prior parameters were used as in the MASALA analysis.

Tables 3 and 4 show the results of the linear model simulation. The fixed-effect parameters β_1 and β_2 are nearly unbiased all estimation methods, with approximately equal MSE and correct nominal coverage. For the intercept parameter β_0, the large bias in the latent class analysis parameters induces bias in the small sample size intercept for the high separation model and all of the low separation models, where the joint modeling methods had reduced bias compared with the two-stage and single imputation methods, and the Bayesian method has both reduced MSE and good coverage except in the smaller sample size. For the latent class α parameters in the well-separated model (Table 3), there is little bias for α_2, because latent class 3 is more clearly distinguished from latent class 1 than latent class 2 is. For α_1, the two-stage and single imputation methods are generally more biased than the joint modeling methods: for the $\text{n} = 500$ and $\text{n} = 200$ case, the expectation–maximization and Bayesian methods perform similarly in terms of bias reduction for α_1; for the $\text{n} = 100$ case, the Bayesian method outperforms the expectation–maximization method, providing essentially unbiased results, whereas the expectation–maximization retains considerable bias. In the poorly separated model (Table 4), the Bayesian approach works well in large sample size, whereas the two-stage and imputation approach are biased for α_2; in the smaller sample sizes, bias and coverage are severely damaged in the two-stage and imputation approach, whereas the Bayesian approach works much better, suffering bias and coverage reduction only in the smallest sample size setting, and then to a much lesser degree than the alternative methods; the expectation–maximization approach results are intermediate between the two-stage and imputation results and the Bayesian approach. For α_1, bias is modest for all methods for $\text{n} = 500$ and $= 200$, and increased for $\text{n} = 100$; coverage is somewhat reduced for all methods except the Bayesian method in the larger sample sizes. In the well-separated model, the increase in variance associated with the correction for measurement error means that there are generally only modest gains if any in MSE for the joint methods over the two-stage or single imputation approach, with the exception of α_1 in the $\text{n} = 100$ case. In the poorly separated model, the reduction in bias

TABLE 3. Simulation Study: Bias, MSE, and Nominal 95% NCP under High Separation Linear Model

	β_0			β_1			β_2			α_1			α_2		
	Bias	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**
n = 500															
Two-stage	−0.015	0.007	5.5	−0.001	0.001	4.5	−0.007	0.008	5.0	0.060	0.015	5.5	0.021	0.012	3.5
Imputation	−0.020	0.007	6.5	−0.001	0.001	3.5	−0.007	0.008	5.0	0.075	0.017	7.5	−0.025	0.012	4.0
EM	−0.002	0.006	7.5	−0.001	0.001	4.5	−0.007	0.008	5.0	0.014	0.013	6.5	0.007	0.012	4.5
Bayesian	−0.004	0.006	5.5	−0.001	0.001	3.0	−0.001	0.008	5.0	0.012	0.012	5.0	0.002	0.012	3.5
n = 200															
Two-stage	−0.001	0.022	9.5	−0.002	0.002	6.5	−0.026	0.020	4.5	0.058	0.034	4.5	−0.007	0.030	4.0
Imputation	−0.008	0.022	9.5	−0.002	0.002	6.5	−0.026	0.020	4.5	0.073	0.035	7.0	0.001	0.030	5.0
EM	0.021	0.016	6.0	−0.002	0.002	6.0	−0.026	0.020	4.5	0.011	0.034	6.5	−0.002	0.031	6.0
Bayesian	0.016	0.017	6.0	−0.001	0.002	6.5	−0.015	0.022	9.0	0.005	0.035	5.0	−0.005	0.030	3.0
n = 100															
Two-stage	−0.033	0.034	5.5	−0.001	0.004	5.0	−0.004	0.037	4.5	0.12	0.090	12	0.027	0.063	5.5
Imputation	−0.038	0.036	5.5	−0.001	0.004	5.0	−0.003	0.037	5.5	0.13	0.094	12	0.031	0.064	4.0
EM	−0.014	0.031	3.5	−0.001	0.004	4.5	−0.005	0.037	5.0	0.067	0.098	12	0.016	0.067	6.0
Bayesian	−0.001	0.027	6.0	0.001	0.003	8.5	−0.009	0.031	5.5	−0.011	0.070	5.5	0.024	0.057	5.5

MSE indicates mean square error; NCP, noncoverage probability.

TABLE 4. Simulation Study: Bias, MSE, and Nominal 95% NCP under Low Separation Linear Model

	β_0			β_1			β_2			α_1			α_2		
	Bias	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**
n = 500															
Two-stage	0.038	0.012	9.0	0.000	0.001	5.5	−0.001	0.008	3.0	0.015	0.018	10	−0.13	0.047	20
Imputation	0.039	0.011	9.0	0.000	0.001	5.0	−0.001	0.008	4.0	0.019	0.018	10	−0.16	0.052	27
EM	0.016	0.010	8.0	0.000	0.001	5.0	0.000	0.008	3.5	0.016	0.022	15	−0.067	0.042	16
Bayesian	0.002	0.008	2.0	0.001	0.001	2.0	0.000	0.007	1.0	0.013	0.017	8.0	0.003	0.018	4.0
n = 200															
Two-stage	0.069	0.038	14	0.005	0.002	6.0	0.003	0.019	4.0	−0.003	0.055	11	−0.24	0.15	25
Imputation	0.065	0.036	13	0.005	0.002	6.0	0.003	0.019	5.0	0.001	0.055	11	−0.24	0.15	26
EM	0.036	0.041	16	0.005	0.002	5.0	0.004	0.020	4.5	0.011	0.074	17	−0.19	0.16	28
Bayesian	0.006	0.023	7.5	0.006	0.002	6.5	0.001	0.019	4.5	0.021	0.048	6.0	−0.038	0.067	7.0
n = 100															
Two-stage	0.087	0.068	10	0.000	0.004	6.0	0.001	0.043	6.0	0.062	0.13	14	−0.35	0.26	34
Imputation	0.094	0.072	12	0.001	0.004	6.0	0.001	0.043	6.0	0.058	0.13	14	−0.37	0.29	34
EM	0.075	0.084	14	0.000	0.004	6.0	−0.001	0.047	7.0	0.057	0.14	18	−0.30	0.27	28
Bayesian	0.033	0.075	11	0.008	0.004	5.5	0.011	0.045	7.0	0.060	0.16	16	−0.16	0.24	14

MSE indicates mean square error; NCP, noncoverage probability.

leads to substantial reductions in MSE for the joint Bayesian method compared with the other methods.

Tables 5 and 6 show the results of the logistic model simulation. The two-stage and single imputation methods are again nearly unbiased for the fixed-effect parameters, with approximately nominal coverage. The joint expectation–maximization and Bayesian methods gave somewhat larger degrees of bias, particularly the Bayesian method, but in all cases it was relatively limited (less than 5%) and had modest if any impact on nominal coverage. MSE actually tended to be reduced for the joint methods, particularly for the Bayesian method, perhaps reflecting the stabilizing impact of the priors not only on the fixed effects but also through the latent class priors (because the covariates and classes are themselves correlated). The joint methods also have approximately correct nominal coverage. For the latent class parameters, the joint methods again afforded improved bias reduction, this time for both α_1 and α_2 parameters; considerable reduction in MSE (on the order of more than 50%) was also achieved in the $n = 200$ and $n = 100$ settings. Despite bias, coverage was approximately

TABLE 5. Simulation Study: Bias, MSE, and Nominal 95% NCP under High Separation Logistic Model

	β_0			β_1			β_2			α_1			α_2		
	Bias	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**
n = 500															
Two-stage	−0.010	0.027	5.0	−0.002	0.003	4.0	−0.031	0.037	4.5	0.071	0.060	6.0	0.021	0.045	5.5
Imputation	−0.018	0.027	5.0	−0.002	0.003	4.5	−0.032	0.037	4.0	0.094	0.064	6.5	−0.029	0.045	4.5
EM	0.014	0.029	3.5	−0.007	0.003	4.5	−0.020	0.038	5.0	−0.036	0.068	3.5	−0.003	0.048	5.0
Bayesian	0.029	0.023	3.0	−0.005	0.003	4.5	−0.059	0.037	6.0	0.014	0.055	5.0	−0.010	0.042	5.0
n = 200															
Two-stage	−0.028	0.079	4.5	−0.006	0.008	3.0	−0.004	0.097	4.0	0.099	0.12	4.0	0.042	0.16	5.0
Imputation	−0.034	0.079	6.0	−0.006	0.008	3.5	−0.004	0.096	4.5	0.11	0.12	4.0	0.049	0.17	4.5
EM	0.014	0.061	4.5	−0.007	0.009	7.5	−0.001	0.082	4.0	−0.021	0.061	1.0	0.015	0.074	1.0
Bayesian	0.059	0.052	4.0	−0.003	0.009	7.5	−0.106	0.084	4.5	0.018	0.078	1.0	0.029	0.091	1.5
n = 100															
Two-stage	0.001	0.16	7.0	−0.003	0.016	4.5	−0.029	0.211	4.0	0.045	0.29	4.5	−0.042	0.26	2.5
Imputation	−0.005	0.16	6.5	−0.002	0.016	4.5	−0.031	0.21	4.0	0.055	0.29	5.0	−0.038	0.26	2.5
EM	−0.011	0.16	7.5	−0.018	0.014	3.0	0.010	0.22	4.5	−0.044	0.086	0.5	−0.004	0.061	0.5
Bayesian	0.094	0.11	3.5	−0.017	0.017	7.0	−0.15	0.18	7.5	−0.031	0.12	0.5	−0.002	0.11	0.5

MSE indicates mean square error; NCP, noncoverage probability.

TABLE 6. Simulation Study: Bias, MSE, and Nominal 95% NCP under Low Separation Logistic Model

	β_0			β_1			β_2			α_1			α_2		
	Bias	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**	**Bias**	**MSE**	**NCP (%)**
n = 500															
Two-stage	0.036	0.030	6.0	−0.001	0.003	7.0	−0.007	0.033	4.0	0.010	0.062	6.5	−0.138	0.083	12
Imputation	0.039	0.029	5.0	−0.001	0.003	7.0	−0.008	0.033	5.0	0.012	0.058	5.0	−0.164	0.092	14
EM	0.054	0.046	6.0	0.005	0.003	5.0	−0.024	0.036	5.0	−0.024	0.044	2.5	−0.062	0.176	18
Bayesian	0.029	0.025	3.5	−0.005	0.004	6.5	−0.028	0.032	4.5	−0.027	0.061	6.5	−0.041	0.076	4.0
n = 200															
Two-stage	0.096	0.096	5.0	0.007	0.008	4.0	−0.027	0.086	4.5	−0.014	0.15	6.0	−0.23	0.26	14
Imputation	0.096	0.098	5.5	0.007	0.008	4.0	−0.028	0.086	5.0	−0.013	0.16	6.0	−0.24	0.26	15
EM	0.080	0.096	5.0	0.009	0.008	4.0	−0.050	0.087	4.5	0.004	0.12	2.5	−0.094	0.40	14
Bayesian	0.062	0.057	3.0	0.001	0.008	3.5	−0.069	0.082	5.5	0.041	0.11	3.0	−0.067	0.15	4.0
n = 100															
Two-stage	0.099	0.17	4.0	−0.010	0.017	5.5	0.009	0.18	3.0	0.040	0.31	5.0	−0.40	0.48	14
Imputation	0.10	0.18	3.0	−0.010	0.017	5.0	0.011	0.18	3.0	0.035	0.31	5.0	−0.42	0.49	13
EM	0.12	0.18	3.0	−0.002	0.018	6.5	−0.015	0.21	5.0	0.034	0.30	2.5	−0.22	0.67	10
Bayesian	0.13	0.13	4.5	−0.027	0.018	4.5	−0.13	0.16	5.5	0.037	0.22	4.0	−0.26	0.28	3.0

MSE indicates mean square error; NCP, noncoverage probability.

nominal for the two-stage and imputation methods; coverage was conservative for the joint expectation–maximization and Bayesian methods for the high separation model; for the low separation model, only the Bayesian model provided approximately correct coverage. For the relatively extreme case of the small sample low separation model, all estimators had substantial bias, although the joint expectation–maximization and Bayesian methods had less bias, and the Bayesian model had much reduced MSE and good coverage despite the bias.

DISCUSSION

Latent class analysis is a powerful tool to summarize large numbers of categorical covariates to understand meaningful structures in such data. It is a natural next step to want to use these classes as predictors in a regression model, where the desired inference focuses on the relationship between the underlying structure in the large number of manifest covariate variables and the outcome. Such applications may be quite common in studies with large numbers of survey questions, as in the MASALA setting, where the focus is not on the association between any one of the specific measures of acculturation and depression, but rather on the collective acculturation strategy that underlies these measures and depression. However, by definition, construction of latent classes involves a degree of measurement error, which can cause bias when they are treated as predictors in a regression model. This issue also occurs in settings such as the use of longitudinal measures to predict survival outcomes (e.g., individual-level trends in prostate-specific antigen measures to predict risk of death from prostate cancer),[35] or the use of individual level variability in longitudinal data settings to predict future health outcomes (e.g., the use of variability to short-term memory tests to predict senility onset).[36] We extend the work in these settings—specifically the use of joint modeling of the latent classes and the regression model—using both an expectation–maximization algorithm and a fully Bayesian method to properly account for the uncertainty in the latent classes when fitting such regression models.

We consider both an application to the development of summary measures of acculturation in predicting depression among older South Asian. Following previous work,[26] we found three latent classes of acculturation: an assimilation class, a separation class, and an intermediate integration class. We found no strong evidence of associations between acculturation class and depression score; however, using a depression indicator of CES-D ≥ 16, we found those in the assimilation class were less likely to experience depression than those in the integration class. In general, the point estimations using joint estimation strategies were larger than those obtained using a two-step method, consistent with our belief that the two-step method will give biased estimates, although there were no substantive differences in the findings between the models. When combined with our results from the simulation study developed out of the MASALA data, it appears that, when all of the classes are well distinguished, use of a two-stage model that ignores measurement error may be warranted. However, as the difference between the classes diminish, measurement error increases and the resulting bias increases, suggesting the need for a joint modeling procedure; our simulation study also suggests that joint models may be more effective when the outcome is binary rather than continuous. To investigate this issue further, the first four columns of Table 7 show, across 200 simulations, the mean, 2.5, and 97.5 percentiles of the average posterior probability of belonging to latent class by true latent class in the high separation simulation, that is, we compute a 3 × 3 table for each high separation simulation, where the j,k th cell, $j = 1,..,3$, $k = 1,..,3$, reports $\frac{\sum_i P(C_i = j \mid Z_{i1},...,Z_{il} \mid C_i = k)}{\sum_i I(C_i = k)}$. We can see from Table 7 that classes 1 and 3 are generally well distinguished from each other, especially when $n = 500$, whereas class 2 is not as well distinguished from class 3 and especially

TABLE 7. Simulation Study: Mean (2.5 Percentile, 97.5 Percentile) Average Posterior Probability of Belonging to Latent Class by True Latent Class

		Posterior Probability					
True Class	**Class 1**	**Class 2**	**Class 3**	**Est. Class**	**Class 1**	**Class 2**	**Class 3**
n = 500				n = 500			
Class 1	0.94	0.042	0.015	Class 1	0.97	0.022	0.009
	(0.88–0.98)	(0.010–0.091)	(0.002–0.034)		(0.94–0.99)	(0.007–0.048)	(0.001–0.021)
Class 2	0.083	0.86	0.055	Class 2	0.046	0.92	0.031
	(0.027–0.17)	(0.75–0.94)	(0.015–0.13)		(0.018–0.082)	(0.87–0.96)	(0.010–0.058)
Class 3	0.030	0.066	0.90	Class 3	0–017	0.026	0.96
	(0.004–0.075)	(0.012–0.17)	(0.78–0.97)		(0.002–0.043)	(0.007–0.050)	(0.92–0.98)
n = 200				n = 200			
Class 1	0.93	0.048	0.018	Class 1	0.98	0.014	0.006
	(0.84–0.98)	(0.006–0.12)	(0.00–0.063)		(0.95–0.99)	(0.001–0.037)	(0.00–0.018)
Class 2	0.087	0.85	0.062	Class 2	0.035	0.95	0.019
	(0.012–0.24)	(0.70–0.96)	(0.002–0.18)		(0.008–0.072)	(0.90–0.98)	(0.003–0.047)
Class 3	0.032	0.062	0.91	Class 3	0.011	0.014	0.98
	(0.001–0.098)	(0.001–0.19)	(0.77–0.99)		(0.00–0.033)	(0.00–0.039)	(0.94–1.00)
n = 100				n = 100			
Class 1	0.92	0.062	0.021	Class 1	0.99	0.007	0.002
	(0.73–1.00)	(0.00–0.24)	(0.00–0.090)		(0.97–1.00)	(0.00–0.025)	(0.00–0.008)
Class 2	0.078	0.86	0.067	Class 2	0.019	0.97	0.007
	(0.001–0.25)	(0.55–0.99)	(0.00–0.24)		(0.001–0.063)	(0.93–0.99)	(0.00–0.025)
Class 3	0.034	0.069	0.90	Class 3	0.005	0.005	0.99
	(0.00–0.14)	(0.00–0.27)	(0.64–1.00)		(0.00–0.021)	(0.00–0.026)	(0.96–1.00)

class 1, with occasional substantial measurement error when sample sizes are small. This would suggest that, as a general rule of thumb, if all of the means of the posterior probabilities of belonging to a latent class are greater than 0.9, measurement error is reduced to the point that a two-stage model is sufficient for using the latent classes as predictors in a second-stage regression model, particularly for continuous outcomes. However, we do not know the actual class memberships, of course, and examining the right four columns of Table 7 shows that, as sample size shrinks, the possibility of overfit increases, leading to misleading results when using this common model fit procedure. Hence, we suggest the "0.9" rule of thumb be applied only in large sample sizes (say in excess of 500–1,000); in small samples some type of cross-validation might be helpful to reduce overfitting.

There are several additional extensions possible from this work. First, because the latent class model likelihoods are invariant to permutations of the label identifiers, it is possible for classes to "switch" during runs of the MCMC chain; this is the so-called "label switching" problem. It appeared that, in both our simulations and application, that labels were stable over MCMC runs; however, this may not always be the case. Stephens[37] develops a postprocessing algorithm to minimize entropy among all possible permutations at each draw that appears to work well to "untangle" the chain in such cases. Next, our current model formulation conditions on covariates for the outcome model, but not for the latent class model, effectively assuming independence between the latent classes and outcome model covariates conditional on the manifest variables. Although this conditional independence assumption might be weaker than it first appears given the richness of the manifest variables (and ease of interpretation), extensions to include covariates to predict in the latent classes and the outcome would be an important next step to pursue. Along the same vein, constraints on the probabilities of the manifest variables that define the latent classes can be imposed to take advantage of ordinal variables, of which we have some in the MASALA data.[38] More recently, latent transition analysis models,[39] an extension of latent class analysis models into a longitudinal setting, have become increasingly popular. The models combine the standard latent class analysis with a hidden Markov model that defines the probability of a subject moving between latent classes at each follow-up period. In its most general form, it defines the classes independently at each time point, although reduced models can be fit. Latent class analysis models imply a fixed number of classes. Determining the number of classes can be a difficult choice, with a variety of model fit measures available.[40–44] In the MASALA model, as in most settings, a mix of model fit measures and substantive science was used to decide on three latent classes in previous work, a choice we carried forward into this manuscript. A more thorough analysis of model fit in these types of joint model settings would be of interest.[45]

REFERENCES

1. Lazarfeld PF. The logical and mathematical foundations of latent structure analysis. In: Stouffer SA, Guttman L, Suchman EA, Lazarfeld PF, Star SA, Clausen JA, eds. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*. Princeton, NY: Princeton University Press; 1950.
2. Clogg CC. Latent class models. In: Arminger G, Clogg CC, Sobel ME, eds. *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum; 1995.
3. Vermunt JK, Magidson J. Latent class analysis. In: *The Sage Encyclopedia of Social Sciences Research Methods*. Thousand Oaks, CA: Sage Publications, Inc.; 2004:549–553.
4. Beran TN, Violato C. Structural equation modeling in medical research: a primer. *BMC Res Notes*. 2010;3:267.
5. VanderWeele TJ. Invited commentary: structural equation models and epidemiologic analysis. *Am J Epidemiol*. 2012;176:608–612.
6. Christ SL, Lee DJ, Lam BL, Zheng DD. Structural equation modeling: a framework for ocular and other medical sciences research. *Ophthalmic Epidemiol*. 2014;21:1–13.
7. Kongsted A, Nielsen AM. Latent class analysis in health research. *J Physiother*. 2017;63:55–58.
8. Jokinen J, Scott JA. Estimating the proportion of pneumonia attributable to pneumococcus in Kenyan adults: latent class analysis. *Epidemiology*. 2010;21:719–725.
9. Hui LL, Schooling CM, Wong MY, Ho LM, Lam TH, Leung GM. Infant growth during the first year of life and subsequent hospitalization to 8 years of age. *Epidemiology*. 2010;21:332–339.
10. Lasry O, Dendukuri N, Marcoux J, Buckeridge DL. Accuracy of administrative health data for surveillance of traumatic brain injury: a Bayesian latent class analysis. *Epidemiology*. 2018;29:876–884.
11. Harlow SD, Karvonen-Guiterrez C, Elliott MR, et al. Does symptom clustering in midlife women foretell healthy aging: the study of women's health across the nation. *Women's Midlife Health*. 2017;3:2.
12. Parada H Jr, Sun X, Tse CK, Olshan AF, Troester MA. Lifestyle patterns and survival following breast cancer in the Carolina Breast Cancer Study. *Epidemiology*. 2019;30:83–92.
13. Berkson J. Are there two regressions? *J Am Stat Assoc*. 1950;45:164–80.
14. Savoca E. Measurement errors in binary regressors: an application to measuring the effects of specific psychiatric diseases on earnings. *Health Services and Outcomes Research Methodology*. 2000;1:149–164.
15. Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*. 1999;55:463–469.
16. Bakk Z, Vermunt JK. Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*. 2016;23:20–31.
17. Bakk Z, Oberski DL, Vermunt JK. Relating latent class membership to continuous distal outcomes: improving the LTB approach and a modified three-step implementation. *Structural Equation Modeling: A Multidisciplinary Journal*. 2016;23:278–289.
18. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, FL: Chapman and Hall/CRC Press; 1995.
19. Bakk Z, Tekle FB, Vermunt JK. Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*. 2013;43:272–311.
20. Collins LM, Lanza ST. *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences (Vol. 718)*. New York: Wiley; 2013.
21. Roeder K, Lynch KG, Nagin DS. Modeling uncertainty in latent class membership: a case study in criminology. *J Am Stat Assoc*. 1999;94:766–776.
22. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc*. 1977;B39:1–38.
23. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC Press; 2013.
24. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*. 2000;10:325–337.
25. Kanaya AM, Kandula N, Herrington D, et al. Mediators of Atherosclerosis in South Asians Living in America (MASALA) study: objectives, methods, and cohort description. *Clin Cardiol*. 2013;36:713–720.

26. Needham BL, Mukherjee B, Bagchi P, et al. Acculturation strategies among South Asian immigrants: the Mediators of Atherosclerosis in South Asians Living in America (MASALA) Study. *J Immigr Minor Health*. 2017;19:373–380.
27. Needham BL, Mukherjee B, Bagchi P, et al. Acculturation strategies and symptoms of depression: the Mediators of Atherosclerosis in South Asians Living in America (MASALA) Study. *J Immigr Minor Health*. 2018;20:792–798.
28. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley; 2013.
29. Radloff LS. The CES-D Scale. A self-report depression scale for research in the general population. *Appl Psychol Meas*. 1977;1:385–401.
30. Linzer DA, Lewis JB. poLCA: an R package for polytomous variable latent class analysis. *J Stat Softw*. 2011;42:1–29.
31. Davison AC, Hinkley DV. *Bootstrap Methods and Their Applications*. Cambridge, UK: Cambridge University Press; 1997.
32. Garrett ES, Zeger SL. Latent class model diagnosis. *Biometrics*. 2000;56:1055–1067.
33. White A, Murphy TB. BayesLCA: an R package for Bayesian latent class analysis. *J Stat Softw*. 2014;61.
34. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219–226.
35. Ye W, Lin X, Taylor JM. Semiparametric modeling of longitudinal measurements and time-to-event data–a two-stage regression calibration approach. *Biometrics*. 2008;64:1238–1246.
36. Elliott MR, Sammel MD, Faul J. Associations between variability of risk factors and health outcomes in longitudinal studies. *Stat Med*. 2012;31:2745–2756.
37. Stephens M. Dealing with label switching in mixture models. *J R Stat Soc*. 2000;B62:795–809.
38. Clogg CC. Some latent structure models for the analysis of Likert-type data. *Social Science Research*. 1979;8:287–301.
39. Lanza ST, Collins LM. A new SAS procedure for latent transition analysis: transitions in dating and sexual risk behavior. *Dev Psychol*. 2008;44:446–456.
40. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974;19:716–723.
41. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2000;7:719–725.
42. Geisser S, Eddy WF. A predictive approach to model selection. *J Am Stat Assoc*. 1979;74:153–160.
43. Speigelhalter D, Best N, Carlin B, Van der Linde A. Bayesian measures of model complexity and fit (with discussion). *J R Statis Soc*. 2003;B64:583–616.
44. Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*. 2010;11:3571–3594.
45. Jiang B, Elliott MR, Sammel MD, Wang N. Bayesian model assessments in evaluating mixtures of longitudinal trajectories and their associations with cross-sectional health outcomes. *Statistics and Its Interface*. 2015;9:183–201.