STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1834–1854
Published online 11 December 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3130

Latent transition analysis: Inference and estimation

Hwan Chung[1,*,†], Stephanie T. Lanza[2] and Eric Loken[3]

[1]*Department of Epidemiology, Michigan State University, MI, U.S.A.*
[2]*The Methodology Center, The Pennsylvania State University, PA, U.S.A.*
[3]*Department of Human Development and Family Studies, The Pennsylvania State University, PA, U.S.A.*

SUMMARY

Parameters for latent transition analysis (LTA) are easily estimated by maximum likelihood (ML) or Bayesian method *via* Markov chain Monte Carlo (MCMC). However, unusual features in the likelihood can cause difficulties in ML and Bayesian inference and estimation, especially with small samples. In this study we explore several problems in drawing inference for LTA in the context of a simulation study and a substance use example. We argue that when conventional ML and Bayesian estimates behave erratically, problems often may be alleviated with a small amount of prior input for LTA with small samples. This paper proposes a dynamic data-dependent prior for LTA with small samples and compares the performance of the estimation methods with the proposed prior in drawing inference. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: latent transition analysis; small samples; MCMC; EM algorithm

1. INTRODUCTION

The analysis of stage-sequential development plays an important role in behavioral and biomedical applications, particularly in the area of substance use prevention and treatment. Using a stage-sequential process, prevention scientists are able to explore risk factors and thus find optimal opportunities for intervening in the substance use onset process. New developments in methods for the analysis of stage-sequential processes include latent transition analysis (LTA). LTA is a latent Markov model, where stage membership at each time is unobserved, but measured with a set of manifest items. In LTA the measurement model at each time point is a latent class model—the associations among manifest items are explained by the underlying categorical latent variable—and

*Correspondence to: Hwan Chung, Department of Epidemiology, Michigan State University, B601 West Fee Hall, East Lansing, MI 48824, U.S.A.
†E-mail: hchung@epi.msu.edu

Contract/grant sponsor: National Institute on Drug Abuse; contract/grant numbers: 5-R03-DA021639, 1-P50-DA10075

Received 25 May 2007
Accepted 4 October 2007

stage-sequential development is summarized in transition probability of latent classes over two consecutive times. LTA has been widely applied to the study of substance use behaviors, including modeling yearly change in risk behavior among injection drug users [1]; exploring risk factors of adolescent substance use onset [2, 3]; assessing transitions between the stages of smoking behavior in the transtheoretical model of behavior change [4]; and exploring change in states of depression throughout adolescence [5].

Maximum likelihood (ML) estimates for LTA are easily estimated by using the expectation–maximization (EM) algorithm [6]. Bayesian inference has also been applied by many researchers *via* Markov chain Monte Carlo (MCMC) [7–13]. However, the LTA likelihood contains some unusual features that can cause difficulties in ML and Bayesian inference [14–17]. Specifically, when the sample size is small and many item-response probabilities are not close to zero or one, the conventional ML and Bayesian estimates behave erratically.

In what follows, we explore difficulties in small-sample inference for LTA by comparing an ML with an MCMC approach in the context of a simulation study. The current study provides a detailed explanation of estimation strategies with possible complications, and their performances are evaluated in terms of estimates and intervals over repeated samples. We show that problems may often be alleviated with a small amount of prior information for LTA with small samples. We then examine stage-sequential patterns of female adolescent substance use and compare the results with various proposed strategies.

2. LATENT TRANSITION MODEL AND ESTIMATION ALGORITHMS

Latent transition models are based on latent class theory, which posits that homogeneous subgroups (i.e. latent classes) of individuals can be identified based on their responses to manifest items. In LTA, the manifest items are measured repeatedly over time to identify latent classes at each occasion, and the probability of transitions over time in latent class membership are estimated. For example, Chung *et al*. [2] used LTA to model stage-sequential patterns of cigarette and alcohol use in order to investigate the relationship between pubertal status and substance use across a range of ages. They identified five classes (No-use, Alcohol only, Cigarettes only, Alcohol and cigarettes, and Cigarettes and drunk) and found that experiencing puberty is related to increased substance use for females in Grades 7–12.

Parameters in the LTA model include class membership probability at time 1, transition probabilities from time 1 to time 2, time 2 to time 3, and so on, and item-response probabilities conditional on latent class. Let $\mathbf{L}=(L_1,\ldots,L_T)$ be the latent class membership from initial time $t=1$ to time T, where $L_t=1,\ldots,L$. Correspondingly, let $\mathbf{Y}_t=(Y_{1t},\ldots,Y_{Mt})$ be a vector of M manifest items measuring the class variable L_t, where each variable Y_{mt} takes values $1,\ldots,r_m$ for $t=1,\ldots,T$. The joint probability that the ith individual belongs to $\mathbf{l}=(l_1,\ldots,l_T)$ and provide item responses $\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT}$ would be

$$P[\mathbf{Y}_1=\mathbf{y}_{i1},\ldots,\mathbf{Y}_T=\mathbf{y}_{iT},\mathbf{L}=\mathbf{l}]=\left[\delta_{l_1}\prod_{t=2}^{T}\tau^{(t)}_{l_t|l_{t-1}}\right]\times\left[\prod_{t=1}^{T}\prod_{m=1}^{M}\prod_{k=1}^{r_m}\rho_{mkt|l_t}^{I(y_{imt}=k)}\right] \tag{1}$$

where $\delta_{l_1}=P[L_1=l_1]$, $\tau^{(t)}_{l_t|l_{t-1}}=P[L_t=l_t|L_{t-1}=l_{t-1}]$, and $\rho_{mkt|l_t}=P[Y_{mt}=k|L_t=l_t]$. We assume that $Y_{1t},\ldots,Y_{Mt}$ are conditionally independent within each class of l_t for $t=1,\ldots,T$. This assumption, called local independence, allows us to draw inference about the latent class variable

DOI: 10.1002/sim

[18]. We also assume that the sequence L_t constitutes a first-order Markov chain for $t=2,\ldots,T$. In (1), only the marginal probability of class membership at initial time $t=1$, δ_{l_1} is estimated; the marginal probabilities of class membership at time t ($\geqslant 2$) are not directly estimated but rather are a function of other parameters. The marginal prevalence of each class at time t ($\geqslant 2$) can be calculated as

$$\delta_{l_t}^{(t)} = P[L_t = l_t] = \sum_{l_1=1}^{L} \cdots \sum_{l_{t-1}=1}^{L} \delta_{l_1} \prod_{j=2}^{t} \tau_{l_j|l_{j-1}}^{(j)}$$

Using (1), the contribution of the ith individual to the likelihood function of $\mathbf{Y}_1,\ldots,\mathbf{Y}_T$ is given by

$$P[\mathbf{Y}_1=\mathbf{y}_{i1},\ldots,\mathbf{Y}_T=\mathbf{y}_{iT}] = \sum_{l_1=1}^{L} \cdots \sum_{l_T=1}^{L} P[\mathbf{Y}_1=\mathbf{y}_{i1},\ldots,\mathbf{Y}_T=\mathbf{y}_{iT},\mathbf{L}=\mathbf{l}] \tag{2}$$

For simplicity, consider a sample of n individuals who responded to M binary items measured at two time periods. We hereafter consider the constrained LTA model where the item-response probabilities (ρ-parameters) are constrained to be equal across time, although an extension to unconstrained LTA is straightforward. For our example, the likelihood function (2) is reduced to

$$P[\mathbf{Y}_1=\mathbf{y}_{i1},\mathbf{Y}_2=\mathbf{y}_{i2}] = \sum_{l_1=1}^{L}\sum_{l_2=1}^{L}\left[\delta_{l_1}\tau_{l_2|l_1}\prod_{t=1}^{2}\prod_{m=1}^{M}\prod_{k=1}^{2}\rho_{mk|l_t}^{I(y_{imt}=k)}\right] \tag{3}$$

where $\tau_{l_2|l_1}=P[L_2=l_2|L_1=l_1]$. Note that $\rho_{mkt|l_t}$ in (1) is reduced to $\rho_{mk|l_t}$. In (3), the free parameters are $\boldsymbol{\theta}=(\boldsymbol{\delta},\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_L,\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_L)$, where $\boldsymbol{\delta}=(\delta_1,\ldots,\delta_{L-1})$, $\boldsymbol{\tau}_l=(\tau_{1|l},\ldots,\tau_{L-1|l})$, and $\boldsymbol{\rho}_l=(\rho_{11|l},\ldots,\rho_{M1|l})$ for $l=1,\ldots,L$.

Under ordinary circumstances, the ML estimates for $\boldsymbol{\theta}$ solve the score equation, $\partial\log\prod_i P[\mathbf{y}_{i1},\mathbf{y}_{i2}]/\partial\boldsymbol{\theta}=0$. Like many finite mixtures, ML estimates for LTA can be estimated using an EM algorithm [6]. For the E-step, we compute the conditional probability that each individual is a member of class l_1 at $t=1$ and class l_2 at $t=2$ given their item responses $\mathbf{y}_i=(\mathbf{y}_{i1},\mathbf{y}_{i2})$ and current estimates $\hat{\boldsymbol{\theta}}$ for the parameters,

$$\hat{\eta}_{i(l_1,l_2)} = P[L_1=l_1,L_2=l_2|\mathbf{y}_{i1},\mathbf{y}_{i2}] = \frac{\delta_{l_1}\tau_{l_2|l_1}\prod_t\prod_m\prod_k\rho_{mk|l_t}^{I(y_{imt}=k)}}{\sum_{l_1}\sum_{l_2}\delta_{l_1}\tau_{l_2|l_1}\prod_t\prod_m\prod_k\rho_{mk|l_t}^{I(y_{imt}=k)}} \tag{4}$$

In the M-step, we update the parameter estimates by

$$\hat{\delta}_{l_1}=\frac{\hat{n}_{l_1}^{(1)}}{n},\quad \hat{\tau}_{l_2|l_1}=\frac{\hat{n}_{(l_1,l_2)}}{\hat{n}_{l_1}^{(1)}},\quad \hat{\rho}_{mk|l}=\frac{\hat{n}_{mk|l}^{(1)}+\hat{n}_{mk|l}^{(2)}}{\hat{n}_{l}^{(1)}+\hat{n}_{l}^{(2)}} \tag{5}$$

where $\hat{n}_{(l_1,l_2)}=\sum_i\hat{\eta}_{i(l_1,l_2)}$, $\hat{n}_{l_1}^{(1)}=\sum_{l_2}\hat{n}_{(l_1,l_2)}$, $\hat{n}_{l_2}^{(2)}=\sum_{l_1}\hat{n}_{(l_1,l_2)}$, $\hat{n}_{mk|l}^{(1)}=\sum_{l_2}\sum_i I(y_{im1}=k)\hat{\eta}_{i(l,l_2)}$, and $\hat{n}_{mk|l}^{(2)}=\sum_{l_1}\sum_i I(y_{im2}=k)\hat{\eta}_{i(l_1,l)}$. Iterating between these two steps produces a sequence of parameter estimates that converges reliably to a local or global maximum of the likelihood function [16, 17, 19].

The challenges for ML inference in small-sample LTA are mainly due to parameters being estimated on the boundary of the parameter space (i.e. zero or one), causing difficulties in obtaining

DOI: 10.1002/sim

adequate standard errors. Although item-response probabilities close to zero or one are highly desirable from a measurement perspective, when some of these parameters are estimated on the boundary, it is impossible to obtain standard errors from the inverted Hessian matrix. To illustrate the problem of a boundary solution, we drew 500 samples of $n=100$ from a two-time, two-class LTA measured by two binary items with parameters $\delta_1=0.5$, $\tau_{1|1}=0.5$, $\tau_{1|2}=0.5$, $\boldsymbol{\rho}_1=(\rho_{11|1},\rho_{21|1})=(0.2,0.4)$, and $\boldsymbol{\rho}_2=(\rho_{11|2},\rho_{21|2})=(0.6,0.8)$, and used the EM algorithm to compute ML estimates. Approximately 45 per cent of the samples in Figure 1(a) and (b) and about 20 per cent of the samples in Figure 1(d) converged to the boundary solution with a criterion of 10^{-7}. The non-normal shape of the sampling distribution of model parameters in Figure 1 suggests that usual large-sample approximations would be inaccurate.

To avoid a boundary solution, Bayesian analysis *via* MCMC has been applied widely over the last decade. The most popular MCMC method for finite mixtures is closely related to EM; it may be viewed either as a Gibbs sampler [20] or a variant of data augmentation [21]. In Bayesian analysis for LTA, we are interested in describing the posterior, $P[\boldsymbol{\theta}|\mathbf{y}_1,\mathbf{y}_2]$. The MCMC algorithm treats the class membership of each individual as missing data, and simulates the augmented posterior $P[\boldsymbol{\theta}|\mathbf{y}_1,\mathbf{y}_2,\mathbf{z}]$ as if class membership were known. Here, the element of $\mathbf{z}=(\mathbf{z}_1,\ldots,\mathbf{z}_n)$, $\mathbf{z}_i$ denotes a two-dimensional array indicating the latent class in which the ith individual belongs, so that $z_{i(l_1,l_2)}\in\{0,1\}$ and $\sum_{l_1}\sum_{l_2} z_{i(l_1,l_2)}=1$. That is, if individual i belongs to class l_1 at time 1 and l_2 at time 2, then $z_{i(l_1,l_2)}$ equals 1 and 0 otherwise. In the first step of the MCMC procedure—the Imputation or I-step—we generate a random draw $z_{i(l_1,l_2)}$ from Multinomial$(1,\eta^*_{i(l_1,l_2)})$ independently for all individuals, where $\eta^*_{i(l_1,l_2)}$ is the posterior probability given in (4) with the observed data $(\mathbf{y}_{i1},\mathbf{y}_{i2})$ and current parameter draws $\boldsymbol{\theta}^*=(\boldsymbol{\delta}^*,\boldsymbol{\tau}^*_1,\ldots,\boldsymbol{\tau}^*_L,\boldsymbol{\rho}^*_1,\ldots,\boldsymbol{\rho}^*_L)$. We then calculate marginal counts $n^*_{(l_1,l_2)}=\sum_i z_{i(l_1,l_2)}$, $n^{*(1)}_{l_1}=\sum_{l_2} n^*_{(l_1,l_2)}$, $n^{*(1)}_{mk|l}=\sum_{l_2}\sum_i I(y_{im1}=k)z_{i(l,l_2)}$, and $n^{*(2)}_{mk|l}=\sum_{l_1}\sum_i I(y_{im2}=k)z_{i(l_1,l)}$. In the second step—the Posterior or P-step—we draw new random values for the parameters from the augmented posterior distribution which regards the latent class membership $z_{i(l_1,l_2)}$ as known. Applying the Jeffreys priors to $\boldsymbol{\delta}$, $\boldsymbol{\tau}_l$, and $\boldsymbol{\rho}_l$, new random values for the parameters are drawn from posterior distributions:

$$\begin{aligned}
\delta^*_1,\ldots,\delta^*_{L-1} &\sim \text{Dirichlet}(n^{*(1)}_1+\tfrac{1}{2},\ldots,n^{*(1)}_L+\tfrac{1}{2})\\
\tau^*_{1|l_1},\ldots,\tau^*_{L-1|l_1} &\sim \text{Dirichlet}(n^*_{(l_1,1)}+\tfrac{1}{2},\ldots,n^*_{(l_1,L)}+\tfrac{1}{2})\\
\rho^*_{m1|l} &\sim \text{Beta}(n^{*(1)}_{m1|l}+n^{*(2)}_{m1|l}+\tfrac{1}{2},n^{*(1)}_{m2|l}+n^{*(2)}_{m2|l}+\tfrac{1}{2})
\end{aligned} \tag{6}$$

for $l_1,l=1,\ldots,L$ and $m=1,\ldots,M$. Repeating this two-step procedure creates a sequence of iterates converging to the stationary joint posterior distribution for $\boldsymbol{\theta}=(\boldsymbol{\delta},\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_L,\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_L)$, given the data. This stream of parameter values (after a suitable burn-in period) is summarized in various ways to produce approximate Bayesian estimates, intervals, tests, etc. [14, 16, 22, 23]. The application of MCMC methods for the LTA model is discussed in [12].

Although the simplicity of EM and MCMC estimation has made the LTA model popular, the likelihood function of an LTA model may have unusual characteristics which can adversely affect inference. For example, there may be certain continuous regions of the parameter space for which the log-likelihood is constant, leading to indeterminacy for some parameters. In the two-class LTA given in (3), suppose that the prevalence of Class 1 is zero at time 1 but then Class 1 emerges at time 2 (i.e. $\delta_1=0$ and $\sum_{l_1}\delta_{l_1}\tau_{1|l_1}>0$). Then the transition probabilities of the individuals in

DOI: 10.1002/sim

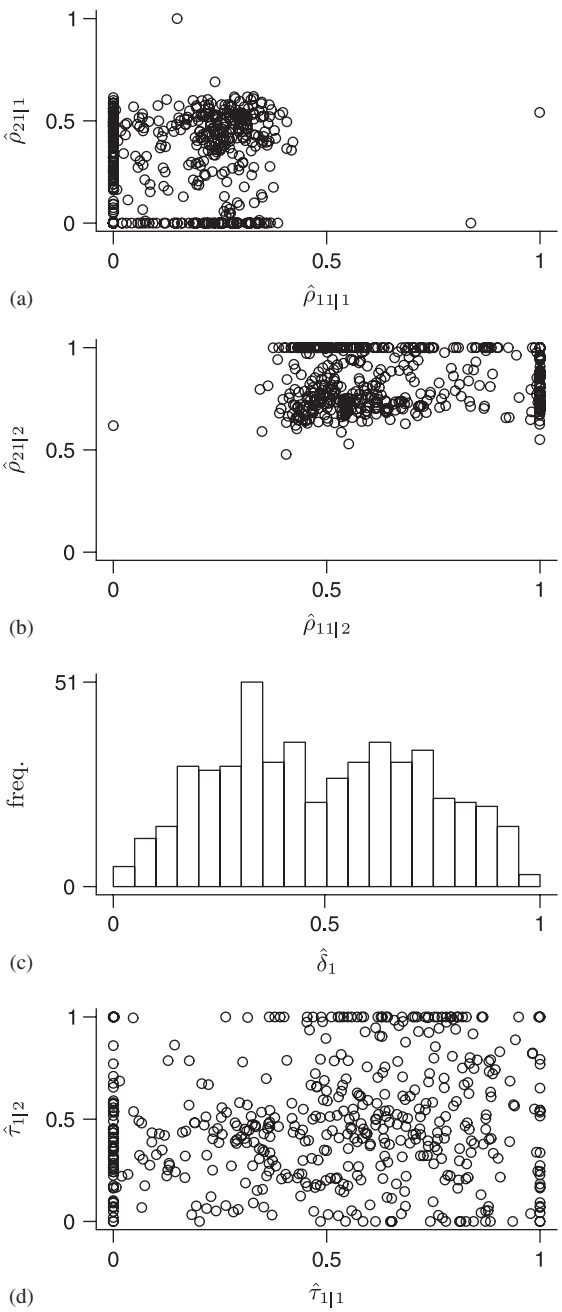

Figure 1. Maximum likelihood estimates for (a) $\boldsymbol{\rho}_1 = (\rho_{11|1}, \rho_{21|1})$, (b) $\boldsymbol{\rho}_2 = (\rho_{11|2}, \rho_{21|2})$, (c) δ_1, and (d) $(\tau_{1|1}, \tau_{2|2})$ from 500 samples with $\boldsymbol{\rho}_1 = (0.2, 0.4)$, $\boldsymbol{\rho}_2 = (0.6, 0.8)$, $\delta_1 = 0.5$, and $\tau_{1|1} = \tau_{1|2} = 0.5$.

DOI: 10.1002/sim

Class 1 at time 1 (i.e. $\tau_{1|1}$ and $\tau_{2|1}$) can take any value from zero to one without any effect on the value of log-likelihood. Because of this unusual geometry, likelihood-ratio test (LRT) statistics for testing hypotheses about the number of classes L are not asymptotically distributed as chi square [16, 24, 25]. In lieu of standard testing methods for choosing an appropriate L, statisticians have turned to penalized likelihood measures such as the Akaike or Bayesian information criteria, to the Lo–Mendell–Rubin LRT [26], to bootstrapping the LRTs [27], and to Bayesian posterior predictive check distributions [9, 28].

3. INFERENCE FOR A SMALL-SAMPLE LTA

Rubin and Schenker [29] applied the Bayesian approach to the standard ML method for a single binomial variable using the Jeffreys prior. They showed that the maximum posterior estimator (MPE) generally outperformed standard maximum likelihood estimator (MLE). For LTA, it is convenient to choose priors that cause $\boldsymbol{\delta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ to be *a posteriori* independent, given $\hat{\boldsymbol{\eta}}=(\hat{\boldsymbol{\eta}}_1,\ldots,\hat{\boldsymbol{\eta}}_n)$, where $\hat{\boldsymbol{\eta}}_i=(\hat{\eta}_{i(1,1)},\ldots,\hat{\eta}_{i(L,L)})$. One way to achieve this is to impose the Dirichlet priors on the joint probabilities of class membership and the item-response probabilities, respectively, as given by

$$P[\boldsymbol{\delta},\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_L]=\prod_{l_1=1}^{L}\prod_{l_2=1}^{L}(\delta_{l_1}\tau_{l_2|l_1})^{\alpha_{(l_1,l_2)}}$$

$$P[\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_L]=\prod_{l=1}^{L}\prod_{m=1}^{M}\prod_{k=1}^{2}\rho_{mk|l}^{\beta_{(m,k)}}$$

Then, the joint posterior for $\boldsymbol{\theta}=(\boldsymbol{\delta},\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_L,\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_L)$ given $\hat{\boldsymbol{\eta}}$ may be expressed as

$$P[\boldsymbol{\theta}|\mathbf{y}_1,\mathbf{y}_2,\hat{\boldsymbol{\eta}}]=\prod_{l_1=1}^{L}\delta_{l_1}^{\hat{n}_{l_1}^{(1)}+\alpha_{l_1}^{(1)}}\prod_{l_2=1}^{L}\tau_{l_2|l_1}^{\hat{n}_{(l_1,l_2)}+\alpha_{(l_1,l_2)}}\prod_{l=1}^{L}\prod_{m=1}^{M}\prod_{k=1}^{2}\rho_{mk|l}^{\hat{n}_{mk|l}^{(1)}+\hat{n}_{mk|l}^{(2)}+\beta_{(m,k)}} \tag{7}$$

where $\alpha_{l_1}^{(1)}=\sum_{l_2}\alpha_{(l_1,l_2)}$. The EM algorithm maximizing (7) is straightforward: the updated parameters for the MPE are obtained by

$$\hat{\delta}_{l_1}=\frac{\hat{n}_{l_1}^{(1)}+\alpha_{l_1}^{(1)}}{n+\alpha},\quad \hat{\tau}_{l_2|l_1}=\frac{\hat{n}_{(l_1,l_2)}+\alpha_{(l_1,l_2)}}{\hat{n}_{l_1}^{(1)}+\alpha_{l_1}^{(1)}},\quad \hat{\rho}_{mk|l}=\frac{\hat{n}_{mk|l}^{(1)}+\hat{n}_{mk|l}^{(2)}+\beta_{(m,k)}}{\hat{n}_{l}^{(1)}+\hat{n}_{l}^{(2)}+\beta_{m}} \tag{8}$$

where $\alpha=\sum_{l_1}\alpha_{l_1}^{(1)}$ and $\beta_m=\sum_k\beta_{(m,k)}$.

The effect of constant hyper-parameters for δ and τ is to smooth the parameter estimates toward equivalent class sizes. For example, the use of $\omega_{l_1}=\alpha_{(l_1,1)}=\cdots=\alpha_{(l_1,L)}$ has a flattening effect on the elements of $\boldsymbol{\tau}_{l_1}=(\tau_{1|l_1},\ldots,\tau_{L|l_1})$ by adding the equivalent of ω_{l_1} observations to each class at time 2 to those who belonged to class l_1 at time 1. For δ-parameters, this prior is equivalent to adding the fictitious $L\times\omega_{l_1}$ observations to class l_1 at time 1. The hyper-parameters $\beta_{(m,k)}$ could possibly depend on the data. For example, we can select $\beta_{(m,k)}\propto\sum_i[I(y_{im1}=k)+I(y_{im2}=k)]/2n$ so that the prior distribution smoothes $\boldsymbol{\rho}$ toward an ML estimate of the raw distribution of the mth item. We drew 500 samples from the LTA model used in Figure 1 and computed maximum posterior (MP)

DOI: 10.1002/sim

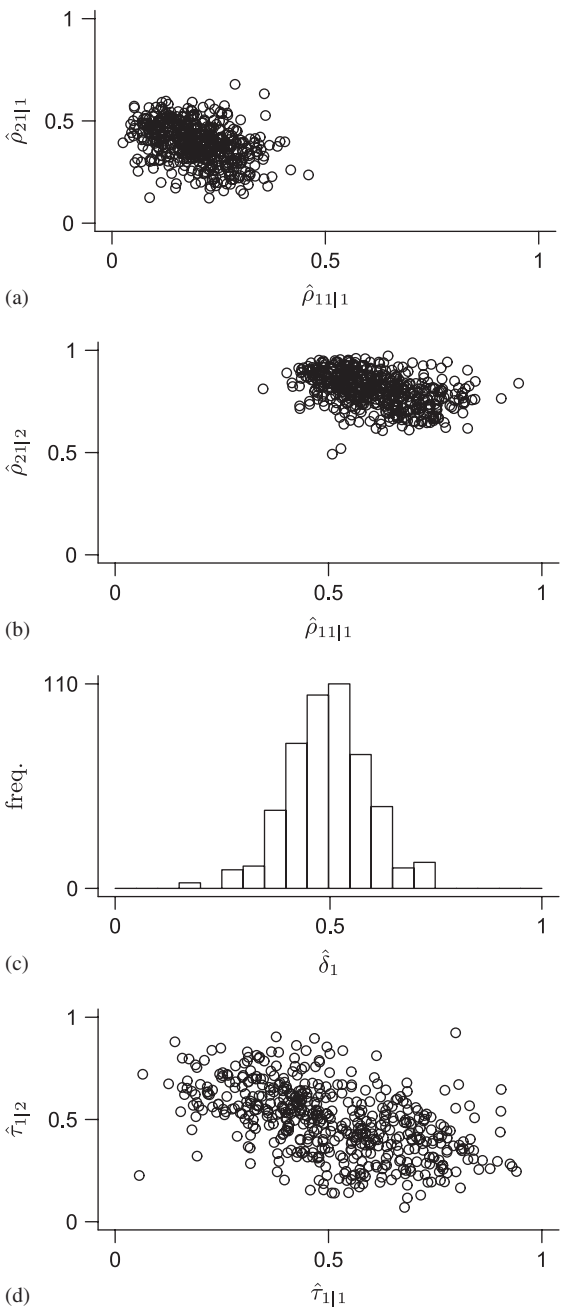

Figure 2. Maximum posterior estimates for (a) $\boldsymbol{\rho}_1 = (\rho_{11|1}, \rho_{21|1})$, (b) $\boldsymbol{\rho}_2 = (\rho_{11|2}, \rho_{21|2})$, (c) δ_1, and (d) $(\tau_{1|1}, \tau_{2|2})$ from 500 samples with $\boldsymbol{\rho}_1 = (0.2, 0.4)$, $\boldsymbol{\rho}_2 = (0.6, 0.8)$, $\delta_1 = 0.5$, and $\tau_{1|1} = \tau_{1|2} = 0.5$.

DOI: 10.1002/sim

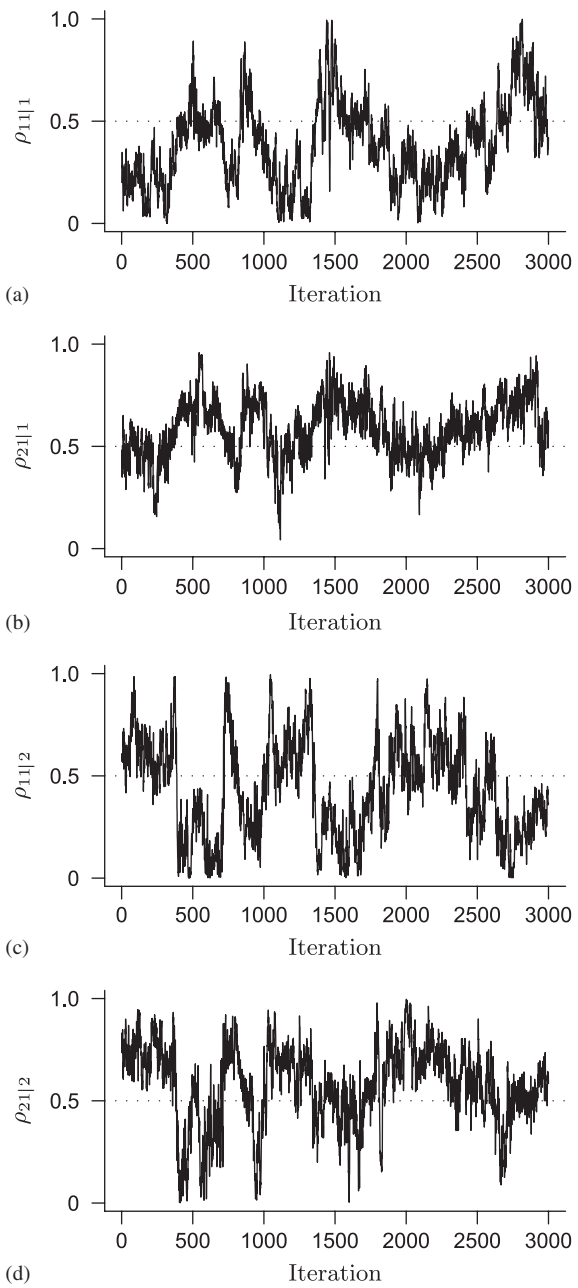

Figure 3. Time-series plots of $\rho_{11|1}$, $\rho_{21|1}$, $\rho_{11|2}$, and $\rho_{21|2}$ over 3000 iterations of MCMC.

DOI: 10.1002/sim

estimates with hyper-parameters $\alpha_{(l_1,l_2)}=1/L$ and $\beta_{(m,k)}=\sum_i[I(y_{im1}=k)+I(y_{im2}=k)]/2n$. Note that this prior is equivalent to adding just one observation to each class at time 1. Plots of the sampling distribution of the MPE are displayed in Figure 2. The boundary solutions evident in Figure 1 are now non-existent.

The likelihood function of the LTA model has multiple equivalent modes which are invariant to permutations of the class labels. When we have two classes in (3) (i.e. $L=2$), $P[\mathbf{y}_{i1},\mathbf{y}_{i2}]$ achieves exactly the same value at $(\delta_1,\tau_{1|1},\tau_{1|2},\boldsymbol{\rho}_1,\boldsymbol{\rho}_2)$ and $(1-\delta_1,1-\tau_{1|2},1-\tau_{1|1},\boldsymbol{\rho}_2,\boldsymbol{\rho}_1)$. This invariant property may cause the most troubling aspect of MCMC for LTA, called label switching. In many applications of MCMC, the analyst runs the algorithm for a 'burn-in' period to eliminate dependence on the starting values, and then saves the subsequent output to estimate meaningful posterior summaries. For certain LTA models, however, the interpretation of long-run averages becomes dubious if the class labels switch during the MCMC run. Label switching can be particularly problematic in applications with smaller samples. To illustrate, we ran MCMC using one sample of $n=100$ from the previous example. We set the prior hyper-parameters to $\frac{1}{2}$ for all parameters, so that all priors become Jeffreys. Time-series plots for $\rho_{11|1}$, $\rho_{11|2}$, $\rho_{21|1}$, and $\rho_{21|2}$ over the first 3000 iterations of MCMC are shown in Figure 3. In Figure 3, note that the order of ρ-parameters reverses many times, requiring the tedious reordering of latent classes before meaningful averages for the parameter can be calculated.

An efficient strategy to handle the label switching is to pre-assign one or more subject's class membership to break the symmetry of the posterior distribution and dampen the posterior density over the nuisance modes [15, 30]. This approach applies constraints to $P[\mathbf{Z}|\boldsymbol{\theta},\mathbf{y}_1,\mathbf{y}_2]$ that result in asymmetric, data-dependent priors for $P[\boldsymbol{\theta}|\mathbf{y}_1,\mathbf{y}_2]$. By pre-classifying a subset of individuals, label switching in the parameter space can be greatly reduced or eliminated. Chung *et al.* [15] showed that the pre-classification method performed well over repeated samples for the exponential finite-mixture models. As the number of classes grows in an LTA model, the choice of which subjects to pre-classify may not be clear. Therefore, we will generalize the pre-classification technique and propose that the choice of which individuals to classify is automated and dynamically adapted throughout the posterior simulation.

4. DYNAMIC DATA-DEPENDENT PRIOR

We face the challenge of determining which individuals to pre-classify in order to achieve good MCMC performance. Although substantive knowledge can often inform this decision, analysts will sometimes need to pre-classify cases blindly. However, some individuals may be particularly informative: consider an individual who has posterior probabilities $\boldsymbol{\eta}_i^*=(\eta^*_{i(1,1)},\eta^*_{i(1,2)},\eta^*_{i(2,1)},\eta^*_{i(2,2)})=(0.97,0.01,0.01,0.01)$ at one iteration during MCMC, implying they have a 97 per cent chance of belonging to the first class at times 1 and 2 given his/her item responses. This individual will 'almost always' be imputed into the first class for both times in the I-step, and should 'almost always' have the same label across iterations. It would be a very strong indication that label switching had occurred if imputed class labels for this individual changed between iterations. A promising strategy is to assign a class label to a particular individual that has a high posterior probability of being in a specific class. A natural criterion is the posterior probability calculated on the basis of the running posterior mean of model parameters $\boldsymbol{\theta}$.

Let $\boldsymbol{\theta}^{(j)}$ represent the generated LTA parameters from the posterior distribution at the jth MCMC iteration, and let $\bar{\boldsymbol{\theta}}^{(N)}=N^{-1}\sum_{j=1}^{N}\boldsymbol{\theta}^{(j)}$ estimate the posterior mean at the Nth iteration. Consider

DOI: 10.1002/sim

the cumulative posterior probabilities at the Nth cycle:

$$\bar{\eta}^{(N)}_{i(l_1,l_2)} = \frac{\bar{\delta}^{(N)}_{l_1}\bar{\tau}^{(N)}_{l_2|l_1}\prod_t\prod_m\prod_k\bar{\rho}^{(N)I(y_{imt}=k)}_{mk|l_t}}{\sum_{l_1}\sum_{l_2}\bar{\delta}^{(N)}_{l_1}\bar{\tau}^{(N)}_{l_2|l_1}\prod_t\prod_m\prod_k\bar{\rho}^{(N)I(y_{imt}=k)}_{mk|l_t}} \tag{9}$$

We can evaluate $\bar{\eta}^{(N)}_{1(l_1,l_2)},\ldots,\bar{\eta}^{(N)}_{n(l_1,l_2)}$ for each combination of classes l_1 and l_2 in order to choose the subject with the largest value. By repeating this procedure independently for all combinations of classes for two-time periods, we can choose L^2 subjects to be identified. We then assign one or more of these L^2 subjects to respective classes with certainty at the Nth cycle. When the latent classes are well differentiated, pre-classifying a small number of subjects out of L^2 subjects may suffice to break the symmetry without introducing serious subjectivity. However, in situations where certain classes are not well differentiated, pre-classifying up to L^2-1 subjects may be preferable. We can pre-classify subjects by a trivial modification of MCMC at the imputation step: if the ith subject is chosen for l_1 and l_2 at cycle N, we deterministically set $z_{i(l_1,l_2)}=1$ at that cycle. This asymmetric penalty to the likelihood, which will be referred to here as a 'dynamic data-dependent prior', tends to dampen the nuisance mode while having very little effect on the mode of interest.

To see how this forced classification translates into an asymmetric prior, we can consider a sequence of draws from the P-step of the data augmentation algorithm. In the previous example (i.e. $n=100$ from a two-time, two-class LTA with parameters $\delta_1=0.5$, $\tau_{1|1}=0.5$, $\tau_{1|2}=0.5$, $(\rho_{11|1},\rho_{21|1})=(0.2,0.4)$, and $(\rho_{11|2},\rho_{21|2})=(0.6,0.8)$), suppose the ith individual has a maximum cumulative posterior probability of belonging to the first class at both times at the Nth cycle. Therefore, we classify him/her to the first class for both times (i.e. $z_{i(1,1)}=1$). Suppose this individual's responses to the items are $\mathbf{y}_i=(y_{i11},y_{i21},y_{i12},y_{i22})=(2,2,2,2)$. In the absence of label switching in the space $\rho_{m1|1}<\rho_{m1|2}$ for $m=1,2$, this is equivalent to draws $\rho_{m1|1}$ from a posterior distribution with a prior Beta$(\frac{1}{2},2+\frac{1}{2})$,

$$\rho^*_{m1|1}\sim\text{Beta}(n^{*(1)}_{m1|1}+n^{*(2)}_{m1|1}+\tfrac{1}{2},n^{**(1)}_{m2|1}+n^{**(2)}_{m2|1}+2+\tfrac{1}{2}) \tag{10}$$

where the imputations have been carried out on the remaining $n-1$ individuals. Compared with (6), marginal counts are partitioned as $n^{**(1)}_{m2|1}=n^{*(1)}_{m2|1}-1$ and $n^{**(2)}_{m2|1}=n^{*(2)}_{m2|1}-1$. Although this may seem like a stronger prior compared with that given in (6), it is actually just a reordering of the terms that (almost) always occurred in the unconstrained MCMC draws without label switching. Therefore, it has very little impact on the posterior distribution in the space $\rho_{m1|1}<\rho_{m1|2}$. All that has occurred is that the fixed individual i is never imputed to be in the second class, whereas he/she was 'almost never' imputed there in the unconstrained case.

In the event of label switching, where the meanings of $\rho_{m1|1}$ and $\rho_{m1|2}$ are reversed, the situation is quite different. The prior for $\rho_{m1|1}$ is now the prior for the class with $\rho_{m1|1}>\rho_{m1|2}$, and in general it would be very rare for responses $\mathbf{y}_i=(2,2,2,2)$ to be imputed to $z_{i(1,1)}=1$. However, the prior distribution Beta$(\frac{1}{2},2+\frac{1}{2})$ does not switch, amounting to a very strong requirement that the class with the smaller ρ-parameter always contains an observation whose responses are $\mathbf{y}_i=(2,2,2,2)$. The impact on the MCMC is to discourage draws from the space $\rho_{m1|1}>\rho_{m1|2}$, thereby restricting the draws to the region of interest and ignoring the symmetric nuisance mode.

This pre-classifying technique can be used to classify more than one individual, although the assumptions on the prior become stronger. For example, suppose we identify the jth individual whose cumulative posterior probability is the largest for the second class at both times, and classify

DOI: 10.1002/sim

him/her into $z_{j(2,2)}=1$ accordingly. Suppose this individual's responses are $\mathbf{y}_j=(1,1,1,1)$. This provides an 'identification' of the second class as the one containing the jth individual. It also stipulates that the ith and jth individuals are never imputed to be in the same class, although draws with both those individuals in the same class are extremely rare in the unconstrained MCMC. With two assigned individuals with responses $(1,1,1,1)$ and $(2,2,2,2)$, times-series plots for parameters over the first 3000 iterations are shown in Figure 4. Comparing these new plots with those of Figure 3, it can be seen that label switching is now almost non-existent. We now evaluate the performance of our procedure over repeated samples.

5. SIMULATION STUDY

The purpose of our simulation study is to investigate the performance of the ML, MP, and Bayesian methods using a dynamic data-dependent prior over repeated samples. These methods represent variations in a solution to handle the problems of estimates on the boundary and label switching during an MCMC run. We drew 500 samples of $n=100$ observations each from the previous example (i.e. two-class LTA with two binary items measured over two-time points with parameters $\delta_1=0.5$, $\tau_{1|1}=0.5$, $\tau_{1|2}=0.5$, $(\rho_{11|1},\rho_{21|1})=(0.2,0.4)$, and $(\rho_{11|2},\rho_{21|2})=(0.6,0.8)$). We assessed the performance of point estimates and nominal 95 per cent confidence intervals for parameters. Four types of estimates were calculated: the standard MLE using the EM algorithm, MPE with hyper-parameters $\alpha_{(c_1,c_2)}=0.5$ and $\beta_{(m,k)}=\sum_i[I(y_{im1}=k)+I(y_{im2}=k)]/2n$ defined in (8), MCMC estimates based on two subjects assigned dynamically across iterations (*DYN-2*), and MCMC estimates based on four subjects assigned dynamically (*DYN-4*). For MLE and MPE, we calculated and inverted the Hessian of the log-likelihood and posterior log-likelihood to obtain interval estimates. For estimates close to zero or one, symmetric interval estimates (estimates plus or minus two standard errors) would not be appropriate as they may stray outside the unit interval. We solved this problem by applying the normal approximation on the logistic scale. For Bayesian methods, we ran MCMC for a 1000 burn-in period plus 2000 iterations, assigning two or four observations to classes for each time period.

Distributions of the Bayesian estimates from the MCMC procedure are shown in Figure 5. Recall that distributions of MLE and MPE were displayed in Figures 1 and 2. Comparing distributions of MLE with those of the Bayesian estimates from *DYN-2* and *DYN-4* in Figure 3, we see that the Bayesian method achieved obvious improvement over MLE.

The average and root-mean-square error (RMSE) of the parameter estimates under each estimation method are given in Table I. All methods except standard ML performed well: the point estimates from MP, *DYN-2*, and *DYN-4* tend to be unbiased, and the values of RMSE are smaller than those from ML. The performance of interval estimates is summarized in Table II, which shows the percentage of intervals that covered their targets and the average interval width. The ML and MP methods fail to produce standard errors for the estimates from 17 and 20 samples out of 500, respectively. The coverages in Table II are based on samples that provide standard errors successfully. ML performed poorly, with coverage lower than the nominal rate of 95 per cent and with wide intervals. Among the MP and Bayesian methods, the interval width of MP tended to be wider and the rate of coverage was low except for τ-parameters. *DYN-2* and *DYN-4*, however, were conservative, exhibiting higher rates of coverage, yet narrower intervals for the parameters than the MP method.

DOI: 10.1002/sim

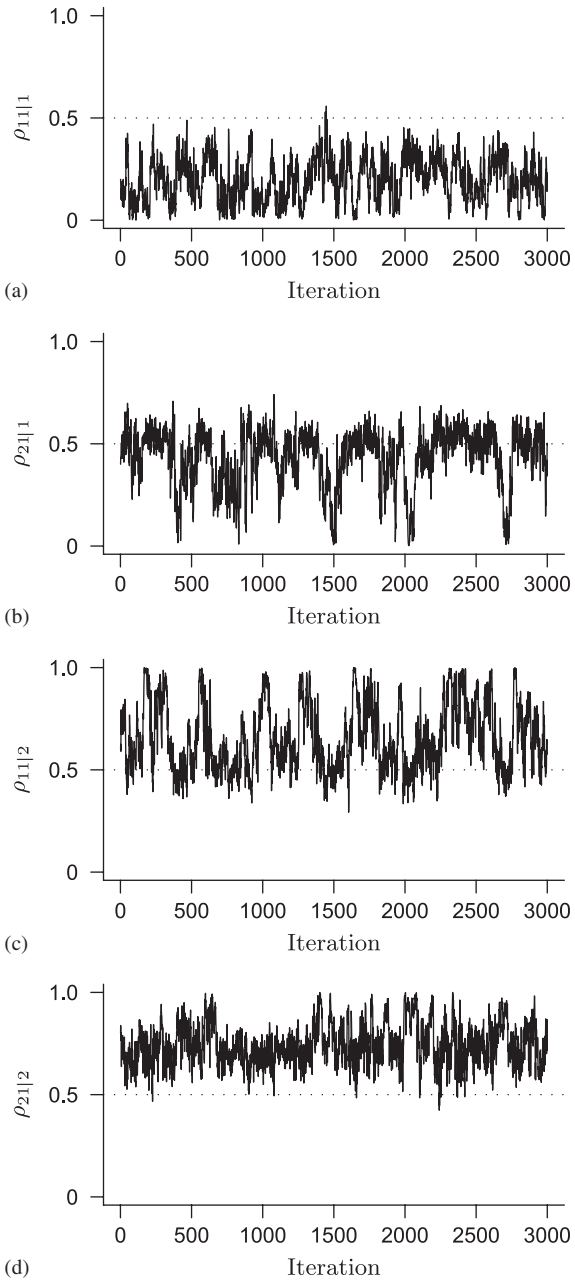

Figure 4. Time-series plots of $\rho_{11|1}$, $\rho_{21|2}$, $\rho_{11|2}$, and $\rho_{21|2}$ over 3000 iterations of modified MCMC.

DOI: 10.1002/sim

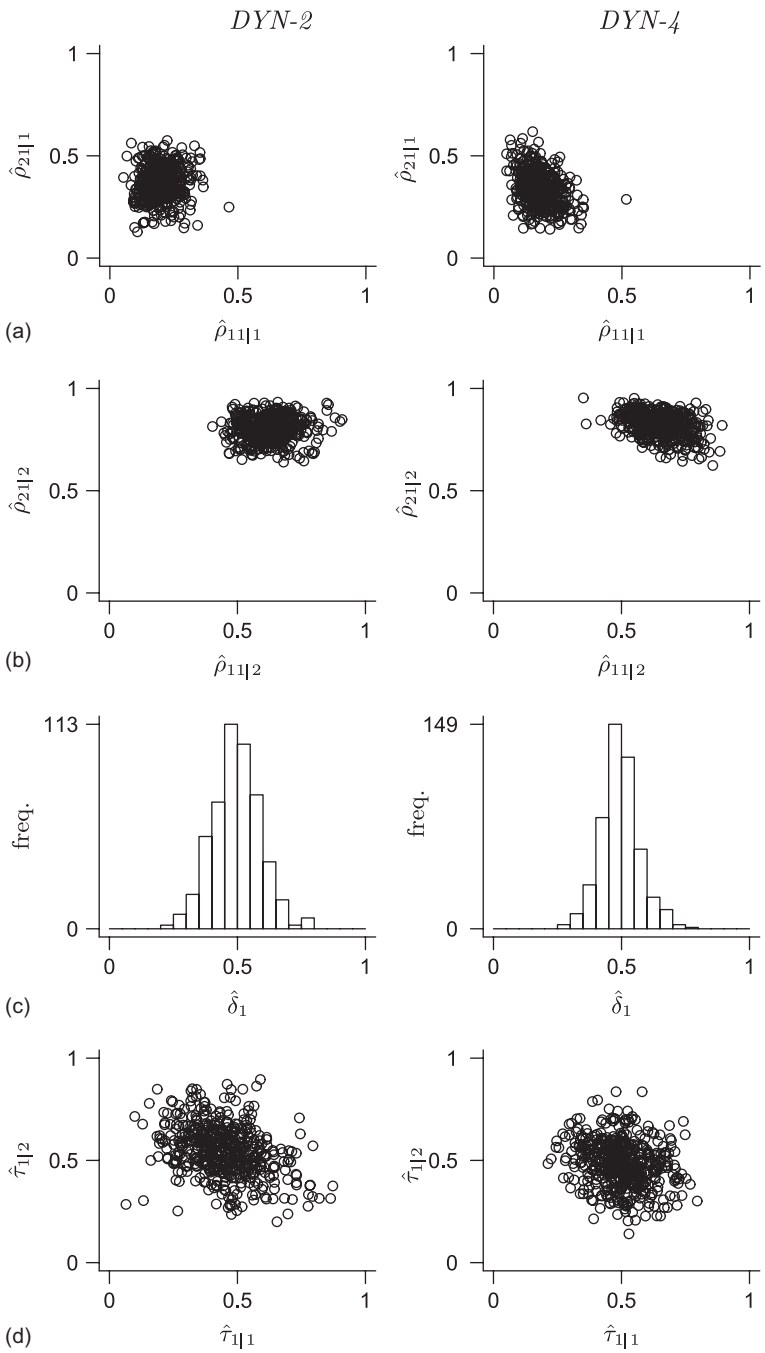

Figure 5. Bayesian estimates for (a) $\boldsymbol{\rho}_1 = (\rho_{11|1}, \rho_{21|1})$, (b) $\boldsymbol{\rho}_2 = (\rho_{11|2}, \rho_{21|2})$, (c) δ_1, and (d) $(\tau_{1|1}, \tau_{2|2})$ from 500 samples with $\boldsymbol{\rho}_1 = (0.2, 0.4)$, $\boldsymbol{\rho}_2 = (0.6, 0.8)$, $\delta_1 = 0.5$, and $\tau_{1|1} = \tau_{1|2} = 0.5$.

DOI: 10.1002/sim

Table I. Average (RMSE) of point estimates over 500 repetitions.

				MCMC	
	True value	ML	MP	*DYN-2*	*DYN-4*
δ_1	0.5	0.495	0.503	0.500	0.500
		(0.237)	(0.094)	(0.094)	(0.075)
$\tau_{1\|1}$	0.5	0.504	0.507	0.447	0.502
		(0.303)	(0.177)	(0.141)	(0.103)
$\tau_{1\|2}$	0.5	0.476	0.491	0.547	0.483
		(0.293)	(0.170)	(0.134)	(0.113)
$\rho_{11\|1}$	0.2	0.172	0.199	0.199	0.180
		(0.143)	(0.077)	(0.057)	(0.060)
$\rho_{21\|1}$	0.4	0.326	0.389	0.366	0.342
		(0.225)	(0.098)	(0.086)	(0.100)
$\rho_{11\|2}$	0.6	0.664	0.609	0.627	0.651
		(0.218)	(0.100)	(0.087)	(0.097)
$\rho_{21\|2}$	0.8	0.833	0.807	0.804	0.821
		(0.135)	(0.082)	(0.057)	(0.060)

Table II. Per cent coverage (average width) of nominal 95 per cent interval estimates over 500 repetitions.

			MCMC	
	ML	MP	*DYN-2*	*DYN-4*
δ_1	78.7	99.8	99.6	99.8
	(0.861)	(0.801)	(0.669)	(0.593)
$\tau_{1\|1}$	71.0	95.6	98.4	99.8
	(0.918)	(0.735)	(0.806)	(0.697)
$\tau_{1\|2}$	74.7	95.8	99.8	99.0
	(0.925)	(0.738)	(0.810)	(0.700)
$\rho_{11\|1}$	93.2	97.5	99.8	99.0
	(0.753)	(0.564)	(0.375)	(0.357)
$\rho_{21\|1}$	88.8	96.2	99.6	99.0
	(0.779)	(0.614)	(0.559)	(0.554)
$\rho_{11\|2}$	87.4	95.6	99.0	99.0
	(0.769)	(0.615)	(0.554)	(0.554)
$\rho_{21\|2}$	94.4	96.7	99.6	99.6
	(0.752)	(0.570)	(0.375)	(0.354)

In summary, in the two-class LTA model, using data-dependent prior information through the MP and Bayesian methods resulted in superior estimation compared with the standard ML. The Bayesian methods with the dynamic allocation based on the MP probabilities perform slightly better than the MP method, but the difference is inconsequential.

DOI: 10.1002/sim

6. AN APPLICATION TO ADOLESCENT SUBSTANCE USE DATA

Our main focus in this paper is to describe the difficulties in inference with small-sample LTA. In order to demonstrate our approach, we draw data for a limited case example from The National Longitudinal Study of Adolescent Health (Add Health) [31]. The Add Health study was conducted to explore the causes of the health-related behaviors of adolescents in Grades 7–12. The first survey, conducted from September 1994 through December 1995, employed a nationally representative sample of 11 796 high-school students. The second wave included the same participants interviewed again between April and August of 1996. The third wave, surveyed from those respondents between August 2001 and April 2002, was designed to collect data useful to examine the transitions from adolescence to young adulthood.

Table III. Estimated class prevalence at wave 1 (δ-parameters) and the probabilities of responding 'yes' to items for each substance use class (ρ-parameters) from the maximum likelihood (ML) and maximum posterior likelihood (MP) methods.

			Observed items			
Method	Class of substance use	Prevalence at wave 1	*AlcUse*	*CigUse*	*5+Drinks*	*Drunk*
ML	1. No use	0.574	0.000	0.000	0.000	0.000
	2. Alcohol	0.125	1.000	0.000	0.044	0.000
	3. Cigarettes	0.076	0.354	1.000	0.000	0.000
	4. Drunk	0.114	1.000	0.087	0.686	0.798
	5. Cigarettes+drunk	0.112	1.000	0.677	0.805	0.873
MP	1. No use	0.561	0.013	0.011	0.001	0.002
	2. Alcohol	0.131	0.924	0.019	0.042	0.037
	3. Cigarettes	0.081	0.355	0.880	0.010	0.012
	4. Drunk	0.098	0.988	0.042	0.712	0.768
	5. Cigarettes+drunk	0.129	0.991	0.932	0.773	0.860

Table IV. Transition probabilities from wave 1 to wave 2 based on the maximum likelihood (ML) and maximum posterior likelihood (MP) methods.

		Class at wave 2				
Method	Class at wave 1	1	2	3	4	5
ML	1. No use	0.594	0.142	0.080	0.116	0.068
	2. Alcohol	0.042	0.344	0.049	0.334	0.231
	3. Cigarettes	0.210	0.000	0.274	0.228	0.288
	4. Drunk	0.051	0.000	0.069	0.570	0.310
	5. Cigarettes+drunk	0.000	0.000	0.163	0.000	0.837
MP	1. No use	0.599	0.148	0.081	0.098	0.074
	2. Alcohol	0.018	0.346	0.055	0.302	0.279
	3. Cigarettes	0.158	0.017	0.303	0.202	0.320
	4. Drunk	0.059	0.024	0.076	0.529	0.311
	5. Cigarettes+drunk	0.010	0.010	0.142	0.022	0.817

DOI: 10.1002/sim

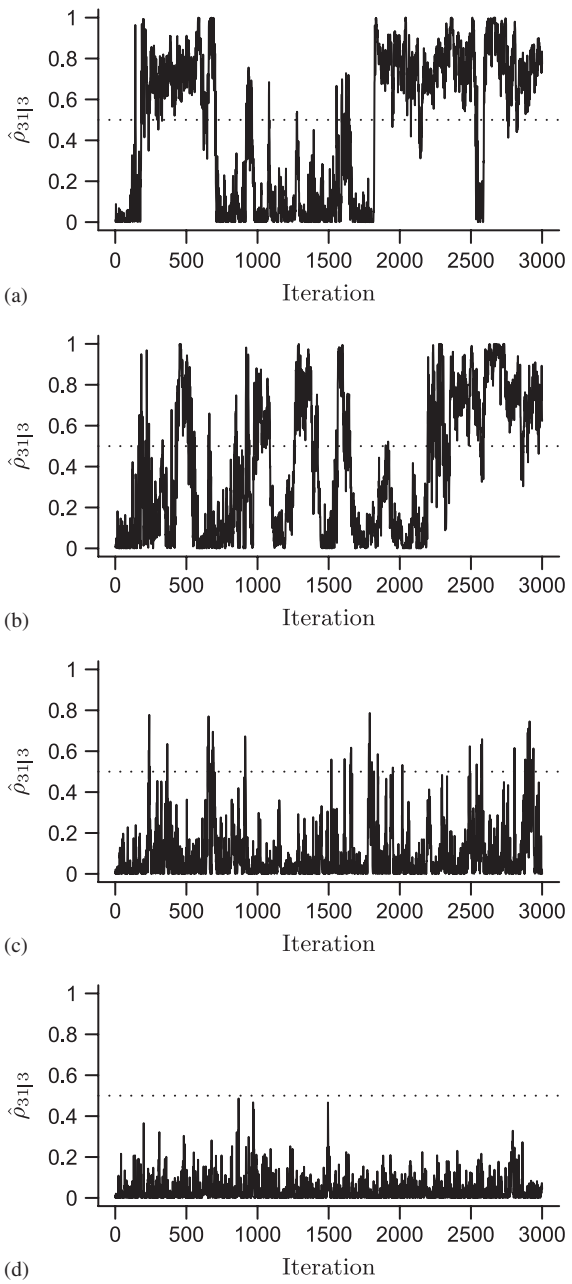

Figure 6. Time-series plots of endorsing *5+Drinks* for Class 3 (i.e. $\rho_{31|3}$) from (a) *MCMC*, (b) *DYN-1*, (c) *DYN-5*, and (d) *DYN-10* over 3000 iterations of MCMC.

DOI: 10.1002/sim

Table V. Summary of the cumulative posterior probabilities for the assigned subjects.

	Modified MCMC		
	DYN-1	*DYN-5*	*DYN-10*
Min.	0.869	0.843	0.755
Max.	0.995	0.989	0.991
Mean	0.942	0.908	0.889

Table VI. Estimated class prevalence at wave 1 (δ-parameters), the probabilities of responding 'yes' to items given substance use class (ρ-parameters), and their 95 per cent confidence intervals based on *DYN-5* and *DYN-10*.

Method	Class	Prevalence at wave 1	Observed items			
			AlcUse	*CigUse*	*5+Drinks*	*Drunk*
DYN-5	1. No use	0.608	0.037	0.063	0.003	0.003
		(0.50, 0.71)	(0.00, 0.13)	(0.00, 0.13)	(0.00, 0.02)	(0.00, 0.02)
	2. Alcohol	0.141	0.917	0.093	0.076	0.112
		(0.07, 0.23)	(0.53, 1.00)	(0.00, 0.29)	(0.00, 0.27)	(0.00, 0.32)
	3. Cigarettes	0.042	0.542	0.836	0.090	0.089
		(0.00, 0.12)	(0.04, 0.97)	(0.40, 1.00)	(0.00, 0.49)	(0.00, 0.46)
	4. Drunk	0.094	0.989	0.229	0.781	0.792
		(0.02, 0.18)	(0.94, 1.00)	(0.00, 0.56)	(0.56, 1.00)	(0.59, 0.99)
	5. Cigarettes+Drunk	0.114	0.988	0.910	0.783	0.884
		(0.04, 0.19)	(0.94, 1.00)	(0.67, 1.00)	(0.58, 0.97)	(0.70, 1.00)
DYN-10	1. No use	0.575	0.016	0.028	0.003	0.003
		(0.48, 0.66)	(0.00, 0.07)	(0.00, 0.08)	(0.00, 0.01)	(0.00, 0.02)
	2. Alcohol	0.153	0.938	0.076	0.076	0.116
		(0.09, 0.23)	(0.72, 1.00)	(0.00, 0.23)	(0.00, 0.22)	(0.00, 0.28)
	3. Cigarettes	0.066	0.305	0.870	0.034	0.030
		(0.02, 0.14)	(0.01, 0.63)	(0.49, 1.00)	(0.00, 0.17)	(0.00, 0.15)
	4. Drunk	0.091	0.989	0.259	0.830	0.817
		(0.03, 0.17)	(0.94, 1.00)	(0.00, 0.63)	(0.60, 1.00)	(0.61, 0.99)
	5. Cigarettes+Drunk	0.115	0.988	0.896	0.725	0.848
		(0.04, 0.18)	(0.94, 1.00)	(0.57, 1.00)	(0.42, 0.89)	(0.63, 0.99)

Early pubertal timing has been identified as a risk factor in relation to early adolescent substance use [3, 32, 33]. We investigate the latent structure of substance use onset in early maturers based on the model presented by Chung *et al*. [2] (i.e. an LTA is specified to include five classes of substance use). They examined stage-sequential patterns of substance use, focusing on measures of cigarette and alcohol use. Using the sample of 3356 females in Grades 7–12 from waves 1 and 2, they found that among 12-year old non-substance users, those who had experienced puberty were approximately three times more likely to advance in substance use than those who had not experienced puberty. The sample used in the current study only includes females aged 12 or 13 who have experienced puberty by wave 1. To identify females who have experienced puberty, we used two items: how developed your breasts are compared with grade school, and how curvy your

DOI: 10.1002/sim

Table VII. Transition probabilities from wave 1 to wave 2 and 95 per cent confidence intervals based on *DYN-5* and *DYN-10*.

Method	Class at wave 1	Class at wave 2				
		1	2	3	4	5
DYN-5	1. No use	0.613 (0.47, 0.73)	0.155 (0.04, 0.31)	0.052 (0.00, 0.15)	0.123 (0.00, 0.15)	0.056 (0.03, 0.24)
	2. Alcohol	0.039 (0.00, 0.17)	0.309 (0.02, 0.59)	0.077 (0.00, 0.31)	0.367 (0.00, 0.58)	0.209 (0.02, 0.76)
	3. Cigarettes	0.141 (0.00, 0.57)	0.104 (0.00, 0.52)	0.269 (0.00, 0.75)	0.207 (0.00, 0.77)	0.278 (0.00, 0.69)
	4. Drunk	0.070 (0.00, 0.29)	0.086 (0.00, 0.39)	0.077 (0.00, 0.31)	0.553 (0.17, 0.90)	0.214 (0.00, 0.57)
	5. Cigarettes + Drunk	0.029 (0.00, 0.15)	0.036 (0.00, 0.16)	0.170 (0.02, 0.42)	0.048 (0.39, 0.24)	0.716 (0.00, 0.93)
DYN-10	1. No use	0.600 (0.49, 0.15)	0.179 (0.09, 0.57)	0.057 (0.00, 0.22)	0.115 (0.03, 0.65)	0.048 (0.00, 0.48)
	2. Alcohol	0.036 (0.00, 0.15)	0.347 (0.14, 0.57)	0.060 (0.00, 0.22)	0.361 (0.10, 0.65)	0.196 (0.00, 0.48)
	3. Cigarettes	0.123 (0.00, 0.39)	0.062 (0.00, 0.31)	0.341 (0.08, 0.71)	0.230 (0.00, 0.64)	0.244 (0.00, 0.64)
	4. Drunk	0.090 (0.01, 0.27)	0.083 (0.00, 0.37)	0.061 (0.00, 0.26)	0.543 (0.17, 0.89)	0.223 (0.00, 0.60)
	5. Cigarettes + Drunk	0.027 (0.00, 0.13)	0.039 (0.00, 0.19)	0.133 (0.01, 0.37)	0.074 (0.00, 0.45)	0.727 (0.31, 0.94)

body is compared with grade school. We included females who reported significant changes in both items at wave 1, resulting in a sample of 202 females.

Four items are used to define adolescent substance use at each year: alcohol use in the past 12 months (*AlcUse*), cigarette use in the past 30 days (*CigUse*), five or more drinks at least once in the past 12 months (*5+Drinks*), and drunkenness in the past 12 months (*Drunk*). All items are rescaled as binary indicators so that possible responses are 1 = yes and 2 = no. Table III gives the parameter estimates from the ML and MP methods for an LTA model with five latent classes of substance use; estimates represent the class prevalence at wave 1 (i.e. δ-parameters) and the probability of responding 'yes' to each of the four items for a given class of substance use (i.e. ρ-parameters). Based on the pattern of these probabilities, the meaning of the five classes can be interpreted and appropriate class labels can be assigned. The point estimates for ML and MP reported in Table III are nearly identical. Table IV gives the transition probabilities of moving from one class of substance use at wave 1 to another class at wave 2 (i.e. τ-parameters). The diagonal values in Table IV are the probabilities of membership in the same class over time. The estimates shown in Tables III and IV are computed by ML and MP methods using EM algorithm, but standard errors are not available because the Hessian matrix cannot be inverted.

The Bayesian method has no difficulty producing intervals, but the class labels may switch during an MCMC run, making the output difficult to interpret. To illustrate, we analyze the same data with small amounts of data-dependent prior information to reduce the label switching. Using a Jeffrey's prior, we perform 2000 iterations after a burn-in period of 1000 cycles. Note that we can identify up to 5^2 subjects who have the largest value in the cumulative posterior probabilities for each

DOI: 10.1002/sim

combination of classes for two-time periods. Four sets of estimates are calculated using a Bayesian approach: the MCMC with no subject assigned (*MCMC*), the MCMC with one subject assigned dynamically across iterations (*DYN-1*), the MCMC with five subjects assigned dynamically across iterations (*DYN-5*), and the MCMC with 10 subjects assigned dynamically across iterations (*DYN-10*). The time-series plots for the probability of reporting *5+Drinks* given membership in Class 3 (i.e. $\rho_{31|3}$) for each MCMC procedure are displayed in Figure 6, showing that the label switching is no longer apparent in *DYN-5* and *DYN-10*. A summary of the assigned individuals' cumulative posterior probabilities is displayed in Table V. For example, during 3000 iterations of *DYN-1*, the average value of the cumulative posterior probabilities for the assigned subjects is close to one (i.e. 0.942), and therefore assigning one subject at each iteration has very little impact on the posterior distribution. However, label switching still occurs because the data-dependent prior given in *DYN-1* is not sufficient to break the symmetry (see Figure 6(b)). By pre-classifying more subjects, the posterior distribution tends to dampen the nuisance mode, but we are introducing more subjectivity (i.e. the average value of the cumulative posterior probabilities is decreasing).

Tables VI and VII give the parameter estimates and their 95 per cent confidence intervals based on *DYN-5* and *DYN-10*. Point estimates from *DYN-5* and *DYN-10* are nearly identical except in *AlcUse* of Class 3. On the basis of the small size of Class 3, it is possible that label switching still occurs in Class 3 for *DYN-5*. However, the point estimates clearly reveal the nature of each of the five classes.

7. DISCUSSION

Many research questions in medical or behavioral science require methods that can classify individuals into pragmatically meaningful groups according to their item-response patterns. The popularity of LTA is increasing because the model can identify population subgroups and their transition probabilities over time. LTA models the structure of subjects' item responses forming discrete classes based on similar item-response profiles. Unlike models for continuous latent variables (e.g. factor analysis and structural equation modeling), this categorical latent variable model does not require further assumptions about the nature of classes. Subjects classified into a particular class have both dimensional and configurational similarities in the distribution of item-response probabilities.

In this study, we explored possible complications in inference for small-sample LTA. In addition, we proposed plausible strategies that can alleviate these problems by using data-dependent priors, and evaluated their performance over repeated samples. We analyzed Add Health data using the ML and the MP methods and demonstrated that Bayesian inference by MCMC may be an attractive alternative. However, new challenges, such as label switching emerge with MCMC, particularly when using a small sample. Recently, Chung *et al*. [15] showed that label switching could be reduced in the Bayesian analysis for exponential mixture models by pre-classifying one or more cases. We have extended their work by developing a dynamic algorithm for selecting subjects to pre-classify in LTA models. We also provided a justification of the procedure to highlight the theoretical contrast with the more traditional approach of relabeling the parameter draws. Although we illustrated our technique in a specific modeling context (LTA), it generalizes immediately to any other mixture case. This automated strategy dynamically assigns class membership of one or more subjects, and thus can be used with mixtures from any family of distributions. The dynamic pre-classification algorithm is easily applicable to the extended version of LTA, where the

DOI: 10.1002/sim

transition probabilities are modeled as a function of covariates. From a computational standpoint, including covariates is a trivial matter, requiring minor modification to the posterior mean of the transition probability (i.e. $\bar{\tau}^{(N)}_{l_2|l_1}$) in (9): the posterior mean of the transition probability can be related to covariates using the logistic link function. The pre-classification technique is easy to implement, and performs at least as well as the standard approach of applying constraints [15]. Other researchers have noted that the choice of constraint to apply can affect inference [14, 23]. As the number of classes increases in LTA, there is difficulty in identifying a unique mode with an appropriate set of constraint compounds.

The technique of pre-classifying a subset of subjects may have some limitations. Pre-classifying k observations into different classes does more than identify k classes—it also assigns these subjects to different classes with certainty. The pre-classification depends on the posterior probabilities of class membership given the parameters, and thus the amount of impact it has on the mode of interest depends on these probabilities. It is expected, however, that the nuisance mode is more adversely affected than the mode of interest. More research is required to compare the pre-classifying technique with traditional constraints across models with different distributions and larger numbers of classes.

Although many substantive interpretations could be pursued in light of the model, we mentioned only a few in this presentation. Note that our exploration was intended to demonstrate a possible solution to the difficulties in small-sample LTA inference and to provide practitioners with a well-worked and plausible example. We provided a limited demonstration using real data to elucidate the model. Our hope is that substantive researchers will be able to identify possible difficulties in estimation and inference for LTA with small samples, and consider using the proposed solution in their research.

REFERENCES

1. Posner SF, Collins LM, Longshore D, Anglin D. The acquisition and maintenance of safer sexual behaviors among injection drug users. *Substance Use and Misuse* 1996; **31**:1995–2015.
2. Chung H, Park Y, Lanza ST. Latent transition analysis with covariates: pubertal timing and substance use behaviors in adolescent females. *Statistics in Medicine* 2005; **24**:2895–2910.
3. Lanza ST, Collins LN. Pubertal timing and the onset of substance use in females during early adolescence. *Prevention Science* 2002; **3**:69–82.
4. Velicer WF, Martin RA, Collins LM. Latent transition analysis for longitudinal data. *Addiction* 1996; **91**: S197–S209.
5. Lanza ST, Flaherty BP, Collins LM. Latent class and latent transition analysis. In *Handbook of Psychology*: *Vol. 2. Research Methods in Psychology*, Schinka JA, Velicer WF (eds). Wiley: Hoboken, NJ, 2003; 663–685.
6. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via EM algorithm (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 1977; **39**:1–38.
7. Chung H, Flaherty BP, Schafer JL. Latent class logistic regression: application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society*, *Series A* 2006; **169**:723–743.
8. Chung H, Walls TA, Park Y. A latent transition model with logistic regression. *Psychometrika* 2007; **72**(3): 413–435.
9. Garrett ES, Zeger SL. Latent class model diagnosis. *Biometrics* 2000; **56**:1055–1067.
10. Garrett ES, Eaton WW, Zeger SL. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Statistics in Medicine* 2002; **21**:1289–1307.
11. Hoijtink H. Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values: applications to educational testing. *Statistica Sinica* 1998; **8**:691–711.
12. Lanza ST, Collins LM, Schafer JL, Flaherty BP. Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Psychological Methods* 2005; **10**:84–100.

DOI: 10.1002/sim

13. Loken E. Using latent class analysis to model temperament types. *Multivariate Behavioral Research* 2004; **39**:625–652.
14. Celeux G, Hurn M, Robert CP. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* 2000; **95**:957–970.
15. Chung H, Loken E, Schafer JL. Difficulties in drawing inferences with finite-mixture models: a simple example with a simple solution. *The American Statistician* 2004; **58**:152–158.
16. McLachlan G, Peel D. *Finite Mixture Models*. Wiley: New York, 2000.
17. Titterington DM, Smith AFM, Makov UE. *Statistical Analysis of Finite Mixture Distributions*. Wiley: New York, 1985.
18. Lazarsfeld PF, Henry NW. *Latent Structure Analysis*. Houghton-Mifflin: Boston, 1968.
19. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: New York, 1987.
20. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**:398–409.
21. Tanner WA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 1987; **82**:528–550.
22. Robert CP. Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice*, Gilks WR, Richardson S, Spiegelhalter DJ (eds). Chapman & Hall: London, 1996; 441–464.
23. Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society*, *Series B* 1997; **59**:731–792.
24. Ghosh JH, Sen PK. On the asymptotic performance of the likelihood ratio statistic for the mixture model and related results. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, vol. 2. Wadsworth: Monterey, 1985; 789–806.
25. Lindsay BG. *Mixture Models*: *Theory*, *Geometry and Applications*. Institute of Mathematical Statistics: Hayward, 1995.
26. Lo Y, Mendell NR, Rubin DB. Testing the number of components in a normal mixture. *Biometrika* 2001; **88**:767–778.
27. McLachlan GJ. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* 1987; **36**:318–324.
28. Rubin DB, Stern HS. Testing in latent class models using a posterior predictive check distribution. In *Latent Variables Analysis*: *Applications for Developmental Research*, von Eye A, Clogg CC (eds). Sage: Thousand Oaks, 1994; 420–438.
29. Rubin DB, Schenker N. Logit-based interval estimation for binomial data. In *Sociological Methodology 1987*, Clogg CC (ed.). American Sociological Association: Washington, DC, 1987; 131–144.
30. Loken E. Multimodality in mixture models and latent trait models. In *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives*: *An Essential Journey with Donald Rubin's Statistical Family*, Gelman A, Meng X (eds). Wiley: New York, 2004; 202–213.
31. Udry JR. *The National Longitudinal Study of Adolescent Health* (*Add Health*), *Waves I & II*, *1994–1996*; *Wave III*, *2001–2002*. Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, 2003.
32. Brooks-Gunn J, Petersen AC, Eichorn D. The study of maturational timing effects in adolescence. *Journal of Youth and Adolescence* 1985; **14**:149–161.
33. Brooks-Gunn J, Reiter EO. The role of pubertal processes. In *At the Threshold*, Feldman S, Elliott GR (eds). Harvard University Press: Cambridge, MA, 1985; 16–53.

DOI: 10.1002/sim