Applied Psychological Measurement
34(7) 483–504

DOI: 10.1177/0146621610362978
http://apm.sagepub.com

Latent Transition Analysis With a Mixture Item Response Theory Measurement Model

Sun-Joo Cho[1], Allan S. Cohen[2], Seock-Ho Kim[2], and Brian Bottge[3]

Abstract

A latent transition analysis (LTA) model was described with a mixture Rasch model (MRM) as the measurement model. Unlike the LTA, which was developed with a latent class measurement model, the LTA-MRM permits within-class variability on the latent variable, making it more useful for measuring treatment effects within latent classes. A simulation study indicated that model recovery using the LTA-MRM was good except for small sample size–short test conditions. A real data application of a mathematics intervention with middle school students indicated that the LTA-MRM clearly detected the intervention effect and also provided a means of helping to better understand the effects compared to a standard multiwave analysis of variance.

Keywords

mixture IRT model, latent transition analysis

Latent transition analysis (LTA) was initially designed with a latent class measurement model (LCM) for investigating stage-sequential change over time (LTA-LCM; Collins & Wugalter, 1992). Methods for modeling stage-sequential latent variables have been useful in a variety of educational and psychological research areas and provide a useful alternative for analysis of some aspects of change. As an example, the LTA-LCM approach can be used to evaluate the effectiveness of an intervention in which differential effectiveness may be likely within different latent classes in the data (Graham, Collins, Wugalter, Chung, & Hansen, 1991). LTA-LCM has not been used extensively with educational or psychological test data, in part because the latent ability in such tests is typically assumed to be continuous, a situation that is not handled by LCM.

This study extends the LTA to include continuous latent variables by incorporating a mixture item response theory (IRT) measurement model (MixIRTM) into the LTA. With this extension of LTA, it is now possible to address potential heterogeneity in change such as occurs in students' response to intervention. B. O. Muthén (2008) introduced a similar modeling framework,

[1]Vanderbilt University, Nashville, Tennessee
[2]University of Georgia, Athens
[3]University of Kentucky, Lexington

Corresponding Author:
Sun-Joo Cho, Vanderbilt University, Peabody Hobbs 213a, 230 Appleton Place, Nashville, TN 37203
Email: sj.cho@vanderbilt.edu

called a hybrid latent transition model, as a longitudinal extension of factor mixture model analysis. The LTA with an IRT measurement model accounts for both latent classes and proficiency with respect to the strategy that is used (i.e., ability within the latent class). Change can occur as a transition from one latent class to another, and as progress or decline along a dimension (e.g., ability) within a latent class. The present article begins with a description of some common approaches to assessment of change using an IRT model and using LTA-LCM. Next, it develops the LTA with an IRT model, incorporating a mixture Rasch model (MRM) into the LTA to form the LTA-MRM. Then the article describes the estimation of LTA-MRM model parameters using Mplus (L. K. Muthén & Muthén, 2006) and presents a simulation study to demonstrate how the LTA-MRM performs in practical measurement situations. Finally, a real data application is presented.

Item Response Theory for Measuring Change

IRT-based longitudinal models (Andersen, 1985; Embretson, 1991; Fischer, 1976, 1989, 1995) are based on modeling of ability as a latent variable. This approach can be very useful for assessing overall effects of treatment or change.

Change in Ability

Fischer's (1976) general linear logistic latent trait model with relaxed assumptions (LLTA) incorporates measurement occasions and change directly in the model. The general LLTA does not fix the number of measurement occasions (or time points), the number of items per time point, or the number of latent dimensions (Fischer, 1989).

Andersen (1985) describes an IRT model developed for analysis of repeated administrations of the same items. The model can be described as follows:

$$P(y_{ijt} = 1|\theta^*_{jt}) = \frac{1}{1 + \exp[-(\theta^*_{jt} - b_i)]}, \quad (1)$$

where θ^*_{jt} is the ability for person j at each time t, and b_i is item difficulty. θ^*_{j1} is taken as the initial ability, and θ^*_{jt} (for $t = 2, \ldots, T$) indicates subsequent change in ability at each time t. Let the ability for person j at each time point be θ^*_{jt} and the change in ability at time t be θ_{jt}. For the first time point, $\theta^*_{j1} = \theta_{j1}$. For $t = 2, \ldots, T$, $\theta^*_{jt} = \theta^*_{j1} + \theta_{jt}$. This approach does not provide a direct estimate of change (Embretson, 1991), although $T-1$ contrasts can be tested simultaneously for calculating ability changes. Below, the LTA is first described with a latent class model (LCM) followed by a description of the LTA with a MixIRTM.

Latent Transition Analysis With Latent Class Model

The LTA is a form of Markov chain model known as latent class analysis (LCA). Markov chain models have been used to analyze individual growth over time and to describe kinds of stagewise development (Langeheine & van de Pol, 2002). LCA is used to detect subgroups that are not directly observed in a population; that is, they are latent. The central difference between LCA and LTA lies in the nature of the latent variable being measured. In this regard, Collins and Wugalter (1992) distinguish between static latent variables and dynamic latent variables. Static latent variables do not change; dynamic latent variables do change, often in systematic ways over time. LTA is intended to measure dynamic latent variables, which are assumed to transition through a series of latent stages over time (Graham et al., 1991).

In a LCM, the probability for observing response pattern, w, is given as

$$P = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{g} \prod_{k=1}^{r_i} \rho_{ik|g}^{I(w_i=k)}, \tag{2}$$

where π_g is the proportion of the population in latent class g, $\rho_{ik|g}^{I(w_i=k)}$ is the probability of response k to item i (i.e., $i = 1, \ldots, g$) having the number of categories r_i for latent class g, and $I(\cdot)$ is defined as 1 if $w_i = k$ and as 0 otherwise.

The LTA-LCM (Collins & Wugalter, 1992) is given as

$$P = \sum_{g_1=1}^{G_1} \cdots \sum_{g_T=1}^{G_T} \pi_{g_1} \prod_{t=2}^{T} \tau_{g_t|g_{t-1}}^{t-1} \left(\prod_{i=1}^{g} \prod_{k=1}^{r_i} \rho_{tik|g_t}^{I(w_{ti}=k)} \right), \tag{3}$$

where π_{g_1} is the proportion of the population in latent class g_1 at time 1, and $\tau_{g_t|g_{t-1}}^{t-1}$ is the transition probability from latent class g_{t-1} at time $t-1$ to latent class g_t at time t. The set of an individual's latent class memberships at one particular time is referred to as a latent status. $\rho_{tik|g_t}^{I(w_{ti}=k)}$ is the probability of response k to item i for the pattern g_t, and $I(\cdot)$ is the indicator function.

When there are three or more occasions of measurement, it is possible for a second-order model to be specified. In a second-order model, transitions between latent statuses are conditional not only on the immediately previous time but also on the time before that as well (Collins & Flaherty, 2002). With three measurement occasions, Time 3 latent status in a second-order model, for example, would be conditional on Time 2 latent status and also on latent status at Time 1.

LTA-LCM is a probabilistic model that uses patterns of categorical responses to estimate examinee status at different time points on the latent variable(s) of interest. It can be used, in other words, to detect latent groups of individuals that differ over time on some set of categorical latent variable(s). LTA-LCM assumes that there is no variability on the latent variable (e.g., on ability) within classes. One concern with LTA-LCM is that as the number of categorical variables (e.g., test items) gets larger, the data matrix becomes increasingly sparse, making it difficult to estimate model parameters (Collins & Wugalter, 1992).

The LTA was originally developed using an unrestricted LCM parameterized in terms of latent proportions and conditional response probabilities. The unrestricted LCMs, however, are not the most useful when the analysis is intended to be for measurement rather than for data reduction (Heinen, 1996). Restricted LCMs (e.g., IRT models) can be obtained by imposing constraints on an LCM. When the intent is to explore potentially useful hypotheses about structure in the data, it is more useful to use restricted LCMs. Below, an LTA is described that incorporates a MixIRTM, thus extending the LTA to the analysis of change on a continuous latent variable.

Latent Transition Analysis With a Mixture Rasch Model

The LCM and IRT in a MixIRTM are both based on the assumption of conditional independence of items. One difference in the two models is that latent ability is categorical in the LCM and continuous in IRT. The result is that the MixIRTM, itself a combination of LCM and IRT, permits within-class variation on the latent variable (i.e., on ability). Members within a latent class formed based on a MixIRTM, in other words, experience the same propensity for a response to each of the items on the test (i.e., the IRT model within a latent class is the same for all members of the class), although examinees in the same class can vary on the latent ability. In the LCM, however, examinees within a latent class are homogeneous in their response to

items and do not vary on latent ability. A second difference is how an increase in test length is handled. When test length increases, the matrix of response patterns becomes increasingly sparse, making it difficult for model parameters to be well estimated. IRT can potentially moderate this sparseness by applying a parametric model (e.g., a cumulative logistic or normal ogive function) with strong assumptions describing the relationship between the response probabilities and the latent ability.

The MixIRTM is typically seen to fit data better than conventional LCM or IRT models (B. Muthén & Asparouhov, 2006). The MixIRTM has been applied to address both the qualitative differences and quantitative differences. Rost (1990, 1997) described an MRM to detect qualitative differences in response characteristics among latent classes of examinees. A mixture linear logistic test model was used to detect random guessing behavior on a multiple-choice test (Mislevy & Verhelst, 1990). Bolt, Cohen, and Wollack (2001) similarly used a mixture nominal response model to investigate individual differences in the selection of response categories in multiple-choice items. An MRM has been used for detecting differential functioning on items (Cho & Cohen, 2010; Cohen & Bolt, 2005), for analyzing testlet effects (Cohen, Cho, & Kim, 2005), and for modeling learning (Wilson, 1989), observed mixtures (von Davier & Yamamoto, 2004), and test speededness (Bolt, Cohen, & Wollack, 2002; Yamamoto & Everson, 1997). Below, a simple MixIRTM, MRM (Rost, 1990), is described and then the article shows how it can be used in an LTA.

Mixture Rasch Model

The probability of a correct response in the logistic form of the MRM can be written as

$$P(y_{ij}=1)=\sum_{g=1}^{G}\pi_g\cdot P(y_{ijg}=1|g,\theta_{jg})=\sum_{g=1}^{G}\pi_g\cdot\frac{1}{1+\exp[-(\theta_{jg}-b_{ig})]}, \tag{4}$$

where g is an index for the latent class, $g=1,\ldots,G, j=1,\ldots,N$ examinees, θ_{jg} is the latent ability of examinee j within class g, π_g is the proportion of examinees for each class, and b_{ig} is the Rasch difficulty parameter of item i in class g.

Latent Transition Model With Mixture Rasch Model

The latent transition model incorporating a MRM as the measurement model (LTA-MRM) can be viewed as a changing-pattern clustering model over time. Patterns are used here to refer to the sequences of movement between latent classes over time. Each cluster follows the model described by Andersen (1985) (Equation 1) except that items can be either the same or different over time. The probability of a correct response is

$$P(y_{ijt}=1)=\sum_{g_1=1}^{G_1}\cdots\sum_{g_T=1}^{G_T}\pi_{g_1}\prod_{t=2}^{T}\tau_{g_t|g_{t-1}}^{(t-1)}\frac{1}{1+\exp[-(\theta^{*}_{jtg_t}-b_{t(i)g_t})]}, \tag{5}$$

where g_t is an index for the latent class, $g=1,\ldots,G, j=1,\ldots,N$ examinees, $\theta^{*}_{jtg_t}$ is the latent ability of a examinee j within the pattern g_t (i.e., a vector of group memberships over time), $b_{t(i)g_t}$ is the Rasch difficulty parameter of item i nested within the form administrated at measurement time t for the pattern g_t, π_{g_1} is the proportion of the population in latent class g_1 at time 1, and $\tau_{g_t|g_{t-1}}^{(t-1)}$ is the transition probability from latent class g_{t-1} at time $t-1$ to latent class g_t at time t. The

LTA-MRM can be extended into a second-order model, similar to that for the LTA-LCM, thereby permitting transitions between latent statuses to be conditional not only on the immediately previous time but on other previous time(s) as well. In B. O. Muthén (2008), the latent class g_t at time t is influenced by both the latent class g_{t-1} at time $t-1$ and potentially influenced by abilities $\theta^*_{jtg_{t-1}}$ at time $t-1$. In the present study, the latent class g_t at time t is influenced only by the latent class g_{t-1} at time $t-1$ for the first-order case.

Transition Proportions

The patterns of movement between latent classes at successive time points are modeled by a stochastic process characterized by a Markov chain that is stationary over time points. Let the state occupied at time t be denoted z_t for $t = 2, \ldots, T$ with G possible group memberships. Then transition probabilities, $p_{g_h h_{t-1}}$, between states are given as

$$p(z_t = g_t | z_{t-1} = g_{t-1}) = p_{g_t g_{t-1}}(t) = p_{g_t g_{t-1}}(t+1) = p_{g_t g_{t-1}} \tag{6}$$

and $\sum_{g_{t-1}} p_{g_t g_{t-1}} = 1$ for the previous group membership, $g_{t-1} = 1, \ldots, G$.

The likelihood of a sequence of states $z_1, z_2, \ldots, z_T$ can be written as

$$p(z_1, z_2, \ldots, z_T) = p(z_1) \prod_{t=2}^{T} p(z_t | z_{t-1}) = p(z_1) \prod_{t=2}^{T} p_{g_t g_{t-1}} = \pi_{g_1} \prod_{t=2}^{T} \tau^{(t-1)}_{g_t | g_{t-1}}. \tag{7}$$

Ability Structure

Multidimensional ability modeling over time in the LTA-MRM is the same as in Andersen (1985), because within each pattern, abilities are time-specific. $\theta^*_{j1g_t}$ can be considered as the initial ability. $\theta^*_{jtg_t}$ for $t = 2, \ldots, T$ involves initial ability and changes at each time t. Let the ability at each time point be $\theta^*_{jtg_t}$ for each pattern and the change in ability at time t be θ_{jtg_t}. For the first time point, $\theta^*_{j1g_t} = \theta_{j1g_t}$. For $t = 2, \ldots, T$, $\theta^*_{jtg_t}$ holds. Although Andersen's (1985) model does not provide a direct estimate of change (i.e., θ_{jtg}), change in ability for the LTA-MRM can be calculated indirectly using ability profiles over time within a given pattern of latent class memberships.

The multidimensional ability structure in an LTA-MRM is given as follows:

$$\theta^*_{jtg_t} \sim MN(\boldsymbol{\mu}_{g_t}, \Sigma_{g_t}), \tag{8}$$

where $\boldsymbol{\mu}_{g_t}$ is the mean vector and Σ_{g_t} is the variance–covariance of the ability dimension across time points in a multivariate normal distribution (MN) for a particular pattern of latent classes. For example, a pattern of 121 indicates a transition from Class 1 at time 1 to Class 2 at time 2 and then to Class 1 at time 3. θ_{jt} is composed of the $\theta^*_{jtg_t}$, which follow $MN(\boldsymbol{\mu}_{g_t}, \Sigma_{g_t})$ with proportions $\pi_{g_1} \prod_{t=2}^{T} \tau^{(t-1)}_{g_t | g_{t-1}}$. $MN(\boldsymbol{\mu}_{g_t}, \Sigma_{g_t})$ is constructed with respect to the transition among the different latent classes for each mixture pattern. In this study, an unrestricted covariance structure for Σ_{g_t} is modeled.

Item Difficulty Structure

Unlike the LTA-LCM (in Equation 3), the LTA-MRM provides sample-independent information about the latent classes at each time point. Item parameter invariance across time points is

assumed to ensure that the number and structure of the latent classes are the same across time in addition to providing a basis for interpretation of the transition probabilities.

Estimation of Latent Structure and Scale Comparability

Item response probabilities in the ability structure described above have a multidimensional structure for each transition pattern. Thus, the probability of a correct response, $P(y_{ijt} = 1|g_t, \theta_{jtg_t})$, can be different when the same items are administered across time points. Two approaches can be considered for item parameter calibration. In the first, class-specific item difficulties are estimated using the data from only the first time point. This is similar to the calibration of item parameter estimates that are subsequently used for other samples from the same population, such as for an item bank. Item difficulties are then held constant across time points. The second approach is joint estimation of item difficulties, abilities, transition probabilities, and group memberships, but holding item parameters invariant across time points. In this case, metric anchoring is obtained by setting the mean and variance of ability for the first time point and the first latent class to 0 and 1, respectively. Item difficulties and abilities are then estimated on this metric. Given this approach to anchoring the metric, it is necessary to place the item difficulty estimates on the same scale, if comparisons are to be made across time points.

The two methods can potentially lead to different latent structures. In the first method, information from the first time point is used to obtain the class-specific item difficulties, under the assumption that the structure does not change across time points. This is referred to as fixed estimation. In the second method, information across time points is used to estimate the latent structure. This is referred to as joint estimation. Both methods were considered in the following simulation study.

Anchor items need to be used to ensure scale comparability among latent classes (von Davier & Yamamoto, 2004). If each latent group of examinees responds to the same set of items, one can think of every item on the scale as being a potential anchor item to be used in estimating an appropriate link (Embretson & Reise, 2000). This is similar to a common-item internal anchor nonequivalent groups linking design, although, in the MRM, class-specific item difficulties as well as group memberships g are estimated simultaneously.

Estimation

The primary modeling framework of the LTA-MRM is a mixture model. There are well-known estimation problems in mixture modeling including identification issues and local solutions (Congdon, 2003; Frühwirth-Schnatter, 2006; McLachlan & Peel, 2000; Vermunt & Magidson, 2005). Below, some of the issues that affect estimation of the LTA-MRM are discussed.

Nonidentification of a model implies that different parameter estimates yield the same log likelihood value. Because the likelihood function is invariant with respect to different permutations of model parameters in the mixture modeling, there is a problem called label switching. Two types of label switching occur with mixture modeling (Cho, Cohen, & Kim, 2006). The first type occurs across iterations within a single chain in Bayesian solution. The second type of label switching can happen for both Bayesian and maximum likelihood estimation (MLE). The second type does not distort the final estimates in an empirical study, but one needs to be aware of it when interpreting results of simulation studies. For example, Class 1 may be generated but the label may be described as Class 2. This kind of label switching can easily be observed in a simulation study because the generating parameters are known.

Maximizing a likelihood yields a solution that provides a local maximum only within a restricted set of parameter values rather than globally over all possible combinations of

Table 1. Labels of Constraints on Ability Structure for Each Transition Pattern for Three Time Periods Used in Both the Simulation Study and the Real Data Study

			Variances (on diagonal) and covariances (Σ_{g_t})		
Mixture-transition pattern		Mean vector (μ_{g_t})	Time 1	Time 2	Time 3
111	Time 1	19	1		
	Time 2	21	7	3	
	Time 3	23	15	11	5
112	Time 1	19	1		
	Time 2	21	7	3	
	Time 3	24	16	12	6
121	Time 1	19	1		
	Time 2	22	8	4	
	Time 3	23	15	13	5
122	Time 1	19	1		
	Time 2	22	8	4	
	Time 3	24	16	14	6
211	Time 1	20	2		
	Time 2	21	9	3	
	Time 3	22	17	11	5
212	Time 1	20	2		
	Time 2	21	9	3	
	Time 3	24	18	12	6
221	Time 1	20	2		
	Time 2	22	10	4	
	Time 3	23	17	13	5
222	Time 1	20	2		
	Time 2	22	10	4	
	Time 3	24	18	14	6

parameter values. As a result, one problem is that the estimation of likelihoods of mixture models, in general, is prone to yielding multiple local maxima (Frühwirth-Schnatter, 2006; McLachlan & Peel, 2000; B. Muthén et al., 2002). In addition, mixture models that are identified in theory can, in certain samples and with certain starting values, be empirically nonidentified (B. Muthén et al., 2002).

The usual method used for checking whether the model is identified or a local solution is obtained is to run the model with multiple different starting values (McLachlan & Peel, 2000). Observing the same log likelihood from multiple sets of initial values increases confidence that the solution is not local.

The computer program Mplus (L. K. Muthén & Muthén, 2006) was used to estimate the LTAMRM in this study. TYPE = MIXTURE causes Mplus to estimate a mixture model. The ESTIMATOR = ML option yields MLE with conventional standard errors and rectangular (specifically, trapezoid) numerical integration as a default (L. K. Muthén & Muthén, 1998-2006).

Maximum likelihood optimization was done in two stages. First, an optimization was carried out for 50 iterations using each of 30 randomly specified sets of starting values generated inside Mplus. Ending values with the highest log likelihoods were used as the starting values in the second stage with the default optimization settings for TYPE = MIXTURE.

Equality constraints are needed to specify the ability structure. Means and variances were set to be equal as group membership was assumed to be the same at each time point. Also,

covariances were set to be equal as group membership was assumed to be the same at adjacent time periods. Three time periods were used in this study. The values in Table 1 are labels indicating which of the different terms in the model were constrained to be equal. There are six unique labels, for example, for the ability means (i.e., 19, 20, 21, 22, 23, and 24). These labels indicate that the means with the same label were constrained to be equal. So for transition patterns 111, 112, 121, and 122, the label 19 indicates that means at Time 1 for these four patterns were constrained to be equal. Similarly, there are six labels for ability variances (i.e., 1, 2, 3, 4, 5, and 6) indicating which variances in Table 1 were constrained to be equal. For example, the variances for the four transition patterns 111, 112, 121, and 122 all have the label 1 for Time 1, indicating that they were constrained to be equal for Time 1.

Simulation Study

This section presents a simulation study designed to illustrate the performance of the LTA-MRM under some practical measurement conditions.

Simulation Design

Two test lengths (14 and 30 items) were simulated along with three sample sizes (100, 1,000, and 3,000 examinees), and two procedures for determining latent structure and metric anchoring over three time periods. A two-group solution was simulated at each time point. Data sparseness in the Shorter Test × Smallest Data Set was considered a potential problem for estimation of covariances among ability estimates. For this combination, therefore, a constraint of 0 on the covariances was imposed as an additional condition.

Generation of Mixture Patterns

The latent classes were generated by manipulating item difficulty profiles between latent groups using the same rationale as in Rost (1990). Each item profile was generated on the same $N(0, 1)$ ability scale so that the quantitative differences (i.e., the differences in ability means between latent classes) not confounded with the qualitative differences as shown in the differences in item profiles. When generated in this way, crossing item profiles reveal qualitative individual differences (De Boeck, Wilson, & Acton, 2005; Rost, 1990). Class 1 was generated for the 14-item test as a group having lower ability, whose members performed well on simple items (Items 7-10). Class 2 was generated as a higher-ability group whose members performed well on more complex items (Items 1-6 and 11-14).

Generating item parameters are shown in Table 2. Assuming item difficulty parameters to be invariant over time, each time period was simulated as having the same class-specific item difficulties. Item difficulties for Class 1 were simulated to be lower than those for Class 2. The simulation study was designed to focus on parameter recovery so anchor items were not considered.

Ability was simulated for the transition patterns shown in Table 3. Σ_{g_t} in Table 3 indicates the standardized variance and covariance matrix for each transition pattern. The assumption was that correlations between abilities decreased with an increase between the simulated time periods for mixture patterns 111 and 222.

Eight possible transition patterns are possible for the two latent classes over the three time points considered in the simulation study. An intervention was simulated between Time 2 and Time 3 by modeling different proportions of each latent class at each time point. As shown in Table 4, 30% of examinees were simulated to be in the high-ability group at Time 2 and 75% at Time 3. In addition, pattern 112 (i.e., initial membership in Class 1, followed by membership

Table 2. Generating Item Difficulties for the Simulation Study

	14-Item test			30-Item test		
Item	Class 1	Class 2	Size of difference	Class 1	Class 2	Size of difference
1	0.4	–2.0	2.4	0.4	–2.0	2.4
2	–1.4	–2.4	1.0	–1.0	–2.4	1.4
3	–1.8	–1.8	0.0	–0.8	–1.8	1.0
4	1.8	1.4	0.4	–1.0	–2.0	1.0
5	2.8	2.0	0.8	–0.8	–1.8	1.0
6	2.0	–1.2	3.2	2.0	1.5	0.5
7	–1.8	1.2	–3.0	0.0	–1.0	1.0
8	–1.7	0.8	–2.5	–2.0	–2.4	0.4
9	–1.8	–1.4	–0.4	–0.5	–1.2	0.7
10	–2.4	2.2	–4.6	0.0	–1.0	1.0
11	–0.2	–1.0	0.8	2.4	2.0	0.4
12	2.4	2.0	0.4	1.4	1.0	0.4
13	–0.2	–1.2	1.0	2.9	2.2	0.7
14	1.9	1.4	0.5	2.4	2.0	0.4
15				–2.8	1.2	–4.0
16				–2.0	2.2	–4.2
17				–2.8	–1.4	–1.4
18				–1.0	2.5	–3.5
19				–1.0	2.0	–3.0
20				–1.8	0.7	–2.5
21				–0.8	–1.0	0.2
22				0.5	0.0	0.5
23				–0.8	–1.2	0.4
24				–2.0	–2.4	0.4
25				–0.5	–1.5	1.0
26				2.0	1.8	0.2
27				2.8	2.4	0.4
28				2.0	1.4	0.6
29				0.0	–2.0	2.0
30				2.8	2.2	0.6

in Class 1 at Time 2, and then membership in Class 2 at Time 3) was simulated to have 50% of the members move from the low-ability group to the high-ability group from Time 2 to Time 3.

Root mean square errors (RMSEs) are given in Table 5 describing recovery of item difficulties, group membership, and transition probabilities for fixed and joint estimation. Recall that in fixed estimation, model parameters were estimated for the first time period and then held fixed across time periods. WINMIRA (von Davier, 2001) was used for convenience for fixed estimation. Item parameters were estimated based on responses from the first time point. For joint estimation, model parameters were estimated jointly across all three time periods.

As shown in Table 5, RMSEs for item difficulties were smaller for joint estimation. Constraining covariance terms to be 0, however, appears to have had no effect on RMSEs for the 14-item $\times$ 100-examinee condition. An increase in test length resulted in slightly smaller RMSEs for the 3,000-examinee condition. RMSEs generally decreased, however, as sample size increased.

The recovery of group membership was assessed as the proportion of the class correctly detected at each time period. Recovery differed little between fixed and joint estimation except for

Table 3. Generating Ability Structure for the Simulation Study

			Variances (on diagonal) and covariances (Σ_g)		
Mixture-transition pattern		Mean vector (μ_{g_t})	Time 1	Time 2	Time 3
111	Time 1	0	1		
	Time 2	0	.65	1	
	Time 3	0	.38	.48	1
112	Time 1	0	1		
	Time 2	0	.65	1	
	Time 3	0	.10	.06	1
121	Time 1	0	1		
	Time 2	0	.02	1	
	Time 3	0	.38	.07	1
122	Time 1	0	1		
	Time 2	0	.02	1	
	Time 3	0	.10	.30	1
211	Time 1	0	1		
	Time 2	0	.08	1	
	Time 3	0	.12	.48	1
212	Time 1	0	1		
	Time 2	0	.08	1	
	Time 3	0	.24	.06	1
221	Time 1	0	1		
	Time 2	0	.49	1	
	Time 3	0	.12	.07	1
222	Time 1	0	1		
	Time 2	0	.49	1	
	Time 3	0	.24	.30	1

the 14-item $\times$ 100-examinee condition with the covariance constraint. Group membership was recovered better with joint estimation. Use of the constraint for the small sample conditions, likewise, yielded smaller RMSEs for joint estimation. As the number of items and examinees increased, the RMSEs decreased. Recovery of group membership was very good in the 1,000- and 3,000-examinee conditions for both fixed and joint estimation.

Recovery of Ability Structure

Table 6 presents average RMSEs for recovery of means, variances, and covariances describing the structure of ability. In the 14-item condition, means were recovered slightly better in the fixed condition, but no differences between fixed and joint estimation were observed in recovery of the means in the 30-item condition.

RMSEs for joint estimation of variances were smaller. Use of the 0 constraint appears to have had no effect on recovery of variances. Recovery of covariances was better for the joint estimation condition, particularly in the larger sample size conditions. Constraining covariances to 0 did appear to improve recovery in the small sample size condition. For all three components of the ability structure (i.e., means, variances, and covariances), the increase in number of items did not affect the recovery. As number of examinees increased, however, the RMSE did become smaller.

Table 4. Generating Values in the Simulation Study

Latent transition probabilities			Group proportions						Transition patterns				
	Time 2					Sample size					Sample size		
Time 1	Latent class			Class	Proportion	100	1000	3000		Proportion	100	1000	3000
Latent class	1	2	Time 1	1	.77	77	770	2310	111	.10	10	100	300
1	.221	.779		2	.23	23	230	690	112	.50	50	500	1500
2	.565	.435	Time 2	1	.7	70	700	2100	121	.05	5	50	150
	Time 3			2	.3	30	300	900	122	.12	12	120	360
Time 2	Latent class		Time 3	1	.25	25	250	750	211	.05	5	50	150
Latent class	1	2		2	.75	75	750	2250	212	.05	5	50	150
1	.786	.214							221	.05	5	50	150
2	.667	.333							222	.08	8	80	240

Conclusions From Simulation Study

Test length did not appear to have much inuence on recovery of model parameters, but recovery of model parameters was generally better in large sample conditions. Recovery for joint estimation likewise tended to be better. RMSEs for the covariance constraint conditions appeared to be smaller for variances and covariances in the fixed estimation condition and slightly smaller for the transition probabilities in both fixed and joint estimation conditions.

Application: Effects of Enhanced Anchored Instruction on Mathematics Achievement in Middle School

Students with learning disabilities (LDs) often have difficulty in identifying and extracting relevant information in typical text-based mathematics problems because of low reading and computational skills (Parmar, Cawley, & Frazita, 1996). Knapp and Turnbull (1990) suggest that these difficulties arise from an overemphasis on teaching low-order skills to low-achieving students at the expense of teaching them more complex concepts. As a result, testing of more complex skills is likely to underestimate what these students might be capable of doing. These students have been shown to perform better in class, however, when instruction is provided using Enhanced Anchored Instruction (EAI) in a technology-rich environment (Bottge, Rueda, Serlin, Hung, & Kwon, 2007). Mathematics problems in the EAI approach are presented by immersing students directly into problem contexts, thus eliminating much of the reading obstacle for many students with LD who have difficulty in both mathematics and reading (see Fuchs & Fuchs, 2002; Geary, 1993; Geary, Hamson, & Hoard, 2000; Muth, 1984). Most students find the EAI problems to be relevant and interesting, contrary to the usual reaction most of them have to solving standard text-based problems (Lesh & Kelly, 2000). Teachers using EAI also place high expectations on their students, a prerequisite for higher academic achievement (Clifford, 1991; Darling-Hammond, 1996).

In this application, the effects of EAI are examined on students' responses to mathematics test questions designed to measure the effects of EAI. Effects of instruction are usually examined at the score level and explained in terms of manifest variables such as membership in a particular demographic group or as a function of ability or achievement level. MixIRTMs also can be used to detect latent groups based on homogeneities in the patterns of responses used by members of a latent group. In this regard, substantial information may be obtained by a careful analysis of

Table 5. RMSEs and Misclassification Rates for Recovery of Generating Parameters for Item Difficulty, Group Proportion, and Transition Probabilities for the Simulation Study

		Item difficulties		Group membership[a]		Transition probabilities	
Number of items	Number of examinees	Fixed	Joint	Fixed	Joint	Fixed	Joint
14	100	.515	.257	.019	.017	.010	.027
	100-C[b]	.515	.285	.025	.014	.049	.014
	1,000	.128	.087	.006	.006	.018	.017
	3,000	.107	.043	.006	.000	.007	.007
30	100	.893	.249	.017	.018	.054	.050
	1,000	.148	.084	.006	.006	.010	.037
	3,000	.074	.052	.000	.000	.006	.005

Note: RMSE = root mean square error.
a. Misclassification rate.
b. Covariance constrained.

patterns of responses by latent groups of individuals to test items (Mislevy & Verhelst, 1990). Previous research has demonstrated that latent groups of students can be identified that differ in the kinds of cognitive strategies they use to answer test questions (Embretson & Reise, 2000; Rost, 1990, 1997).

In this application, an LTA-MRM was used in the assessment of effects of a multiwave experiment of an EAI instructional treatment (Bottge, Heinrichs, Chan, Mehta, & Watson, 2003) on mathematics achievement of LD and non-LD middle school students. If EAI instruction is effective at improving LD students' mathematics test performance on items requiring complex skills, then they should be better able to correctly answer these items following EAI instruction than before. To investigate this conjecture, the present study examined students' responses to performance-based test questions requiring use of multiple mathematical skills for constructing and describing solutions to real problems involved in building a skateboard ramp. The LTA-MRM was used to determine whether particular sets of patterns emerged that may be due to changes in the differential use of response strategies for answering these performance test items. Results are presented for the EAI instructional treatment Fraction of the Cost (FOC; Bottge et al., 2003) described below.

Methods for EAI Application

FOC instruction. The FOC video-based anchor (Bottge et al., 2003) immerses students in building a skateboard ramp. The problem is introduced in an 8-minute video-based anchor called Fraction of the Cost. This video stars three students from a middle school in a small midwestern town. The video, available in both Spanish and English, was filmed at a local skateboarding store and in the garage and backyard of a local home in the community. Students are asked to learn the mathematics needed to solve multiple real-world problem-based performance tasks as they work to build the ramp. The FOC problem requires them to read schematic plans to determine how much wood they will need; how much the materials, including the wood, the screws, etc. will cost; and what size pieces of wood they will need. Then they are asked to determine the best way to cut each of the pieces without wasting wood, and then to actually build the ramp.

Two mathematics teachers (MT1 and MT2) each taught three 90-minute blocks per day. MT1 taught an inclusive class, a high-achieving class (prealgebra), and a class of average-achieving (typical) students. Pretest scores on the Iowa Tests of Basic Skills (ITBS [Form A]; University of

Table 6. RMSEs for Recovery of Ability Structure by Type of Estimation for the Simulation Study

		Means		Variances		Covariances	
Number of items	Number of examinees	Fixed	Joint	Fixed	Joint	Fixed	Joint
14	100	0.281	0.370	0.763	0.326	0.410	0.273
	100-C[a]	0.241	0.307	0.587	0.399	0.290	0.246
	1,000	0.039	0.045	0.082	0.079	0.246	0.243
	3,000	0.018	0.014	0.041	0.129	0.165	0.128
30	100	0.276	0.280	0.346	0.306	0.316	0.337
	1,000	0.053	0.034	0.096	0.060	0.094	0.081
	3,000	0.023	0.016	0.018	0.016	0.035	0.029

Note: RMSE = root mean square error.
a. Covariance constrained.

Iowa, 2001) confirmed the difference between MT1's three classes in computation and in problem solving. MT2 taught three typical classes. Pretest ITBS scores confirmed there were no differences between MT2's typical classes in either computation or in problem solving.

FOC mathematics test. A test to measure effects of the FOC instruction was constructed for assessing knowledge of mathematics contained in the standards recommended for students in Grades 6 to 8 by the National Council of Teachers of Mathematics (specifically, numbers and operations, measurement, problem solving, communication, and representation). The FOC test used in this study was a 23-item test composed of performance tasks assessing students' abilities to interpret a three-dimensional schematic plan, measure lengths of building materials in feet and inches, estimate and compute combinations using whole numbers and fractions, and interpret and record data in tables. Responses to 20 of the 23 items consisted of constructing short answers that were scored as either correct or incorrect. Six of these 20 items had very low or zero proportions of correct responses for the first two time points. Five of these six were problem-solving items. These six items were removed from the analysis, leaving 14 dichotomously scored items to be analyzed for this application.

Data sources. Students were drawn from six mathematics classrooms in a small Midwestern school district. The sample consisted of 50 males and 59 females, all in the seventh grade. There were 23 high-achieving (prealgebra) students, 69 average-achieving (typical) students, and 17 low-achieving students. Nine students had diagnosed LDs and were classified by the school as low-achieving. One student with LD was not classified by the school as low achieving. The remainder of the students in the sample had no LD.

Multiwave study design. The study spanned 7 months of the regular school year, from October to April. The FOC tests administered at three time points during the course of the year were analyzed in this study. Time point 1 (Pretest 1) was administered in October to 109 students. The FOC test was administered again at Time 2 (Pretest 2) in November to 107 of the 109 students (immediately following another instructional intervention, Kim's Komet). The FOC test was administered again to 109 students at Time 3 (Posttest) in April immediately following FOC instruction. (The Kim's Komet instructional intervention took place over a 3-week period in November of the school year following the FOC pretest. The mathematics taught in the Kim's Komet treatment was designed to not interact with that taught in the FOC treatment.)

LTA-MRM analysis. An exploratory analysis was done to fit an MRM to the data for each time period using the computer program WINMIRA (von Davier, 2001). The number of latent classes at each time period was determined by selecting the model that provided the best fit to the data. This was then followed by joint estimation of model parameters given the best-fitting model at

Table 7. Cognitive Skills Required for Solving Each Item for the Real Data Study

	Cognitive skill		
Item	Number and operation	Measurement	Representation
1	1	0	0
2	1	0	0
3	1	1	0
4	1	1	0
5	1	1	0
6	1	1	0
7	1	1	0
8	1	1	0
9	0	1	0
10	1	1	1
11	1	1	1
12	1	0	0
13	1	0	0
14	1	0	0

each time point. The Bayesian information coefficient (BIC) has been shown to perform well for selection of the correct number of latent classes in simulated data for MRM (Li, Cohen, Kim, & Cho, 2009). According to the BIC, a two-group model was the best fit for all three time points.

In this application, only those items on the FOC test that required more than one mathematics skill or operation to arrive at a correct answer were analyzed. These items were more complex than those that only required a single mathematics skill. Table 7 presents a matrix indicating the particular mathematics skills that are required to solve each item. As can be seen in Table 7, Items 3, 4, 5, 6, 7, 8, 10, and 11 required two or more different skills for a correct answer and so were considered to be complex items. Items 1, 2, 9, 12, 13, and 14 were not considered complex items as they required only a single skill for a correct answer. Instead, these items were evaluated as possible anchor items to anchor the metric across latent classes.

Anchor items were used to ensure scale comparability among parameters in the two latent classes. The equality of the estimated item parameters across the latent classes was determined by separately constraining the item difficulties for one item at a time over the two latent classes. The likelihood for this model was compared to that for a model in which all item parameter constraints were relaxed so that estimates for all items could potentially be different for the two latent classes. A likelihood ratio test was used to compare the -2 log likelihoods for the one-item-constrained and all-items-fully-relaxed models. In this way, each item was examined separately based on data from the first time point to determine the possibility it was class-invariant. No change in the -2 log likelihoods was observed between the all-items-fully-relaxed model and the one-item-constrained models for Items 2, 12, and 13. Based on this result, these three items were used as anchor items. Furthermore, the mean of ability for the first time point and the first class was set to be 0 for the model identification. Joint estimation was then used to estimate the remaining item difficulties over the three time periods in Mplus (L. K. Muthén & B. Muthén, 2006).

Data were sparse in some response patterns, as the FOC sample size was small. This, in turn, resulted in difficulty in estimating covariance terms for the ability structure. The LTA-MRM was first fit by setting all covariances as unconstrained and, therefore, estimated freely. Transition probabilities estimated between time periods were quite small and appeared to be poorly

estimated. As a result, the following ability covariances terms were fixed at 0: Between Class 2 at Time 1 and Class 1 at Time 2, between Class 1 at Time 2 and Class 1 at Time 3, and between Class 2 at Time 2 and Class 1 at Time 3.

First- and second-order LTA analyses were then estimated, and a first-order LTA solution was selected based on BIC as better fitting the data. LTA-MRM results for 50 starting values in Mplus indicated that the highest log likelihood was replicated in the five final stage solutions indicating that a local solution probably was not reached and the model was identified for the first-order joint estimation.

Results for the Application

Class characteristics. Class characteristics were investigated with respect to item difficulty, ability estimates, and association between latent and manifest group memberships. In general, members of Class 1 had difficulty providing correct answers to items that dealt with measurement and with questions that required multiple skills. Further analysis of their responses indicated that compared to members of Class 2, members of Class 1 tended to give answers that failed to account for measures on the schematics. Below, it is described in more detail what these differences were.

Item difficulties for the LTA-MRM were estimated using joint estimation. This assumed invariance of item parameters across the three time periods. Plotting item difficulties for different latent classes can sometimes provide a visual means of determining strengths and weaknesses of each group. Figure 1 presents such a plot for the two latent classes in the application: Item 1 and Items 3 to 10 were clearly more difficult for members of Class 1, and Items 11 and 14 were more difficult for members of Class 2. As mentioned earlier (and as shown in Table 7), Items 3 through 8, 10, and 11 required two or more different skills for a correct answer so were considered to be complex items. Members in Class 2 were more likely to solve these items than members in Class 1. Items 3 to 8 ask students to interpret a schematic plan, list the number and lengths of wood required for building the skateboard ramp, and then convert the measurements from feet-and-inches to inches. For Items 10 and 11, students must figure out and show how to cut 2×4s to waste as little wood as possible. This task is not straightforward because it requires students to use the most economical combinations of wood from the garage. Students need to read a tape measure to measure the wood (i.e., measurement), figure out how this wood can be used based on their interpretation of the schematic plan (i.e., representation), and then compute the combinations (i.e., numbers and operations). Items 1 and 2 ask students to figure out how much money one of the friends shown in the video can contribute to the skateboard ramp project. In the video, one boy states he can spend 10% of the $210 in savings. Looking at a bank statement, his friend states that he must keep $50 of his lawn mowing money in the bank. Items 12 to 14 were not considered complex, although answers to the problems depended on a correct interpretation of the overall FOC problem and answers given to previous items on the test. For example, Item 14 asked students to figure out the total cost of building the ramp. The math skill required is simply adding the dollar amounts of the items (e.g., 2×4s, lumber, plywood, screws). However, the correct cost is dependent on their interpretation of the schematic plan, calculating the economical use of wood, and how much new wood they need to purchase.

Class 2 clearly was higher in ability than Class 1 over all three time points. There were 64 students in latent Class 1 at Time 1 ($M = 0$, $SD = 0.93$), 42 in Class 1 at Time 2 ($M = 0.45$, $SD = 0.78$), and 10 in Class 1 at Time 3 ($M = 0.95$, $SD = 0.47$). In Class 2, there were 45 students at Time 1 ($M = 0.42$, $SD = 0.73$), 67 at Time 2 ($M = 1.01$, $SD = 0.86$), and 99 at Time 3 ($M = 2.02$, $SD = 1.14$).

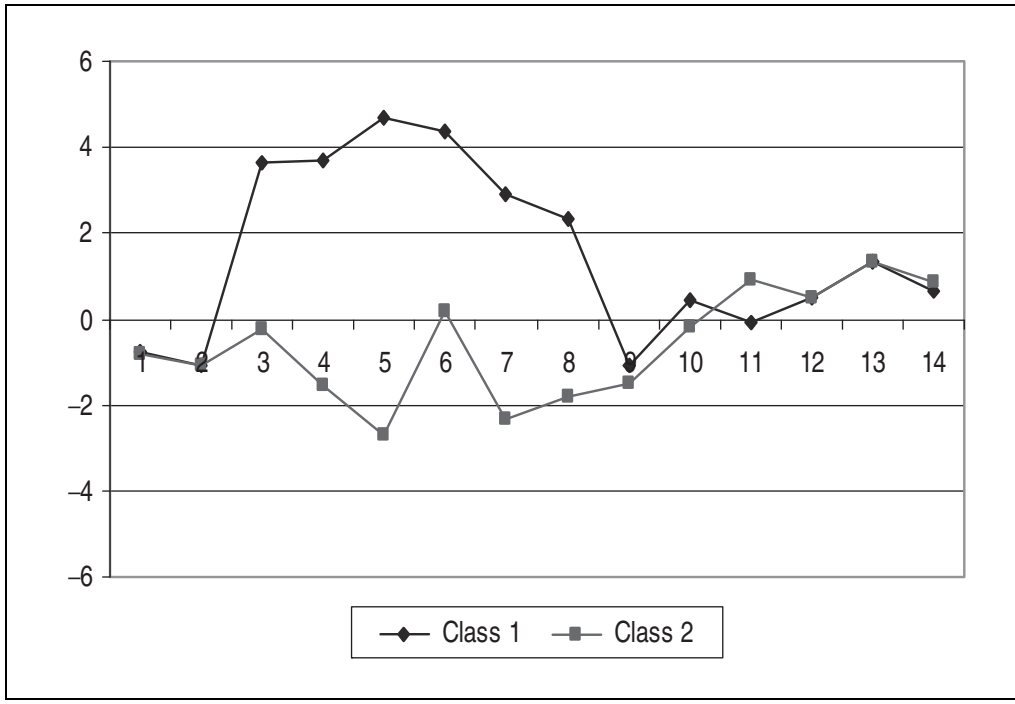

Figure 1. Item difficulty profiles for the application.

Fourteen of the 17 low-achieving students were in Class 1 on Pretest 1, 11 were in Class 1 on Pretest 2, and no low-achieving students were in Class 1 on posttest. Eight of the nine LD students were in Class 1 at Pretest 1 and Pretest 2. All students with LD were in Class 2 at Time 3. That is, all students with LD had moved into the high-ability class following FOC instruction. A chi-square on latent class by achievement level and latent class by LD status indicated that both achievement level and LD status were related to latent class membership for the first two time periods (i.e., before FOC instruction). There was no association for these variables, however, following FOC instruction. Gender also was not related to class membership at any time point.

Proportions and counts for each transition pattern. Proportions for each transition pattern are shown in Table 8. The three most frequently observed transition patterns were 112, 122, and 222: 30 students had Pattern 112, 27 students had Pattern 122, and 36 students had Pattern 222. There are some transition patterns like 121, 211, and 221 that are less plausible given the design of the study. These transition patterns have very low frequencies and possibly represent classification errors given the sample sizes in an empirical study (109 examinees and 14 items). Results from the simulation study suggest that the misclassification rate for group membership was .017 for the joint estimation case for a 14-item test and a 100-examinee sample (see Table 5).

Eight of the nine students with LD had Pattern 112, and one student with LD had Pattern 222. According to the LTA-LCM results, in other words, the one LD student who was not low achieving was in the high-ability group at Pretest 1 and Pretest 2, and also in the high-ability group at posttest.

Two omnibus hypotheses were tested using a split-plot design based on the transition pattern variable: (a) Were the changes in ability over time the same across transition patterns (i.e., was there an interaction effect between time and latent class)? and (b) For all transition patterns, was

Table 8. Class Counts and Proportions for Each Transition Pattern for the Real Data Study

Pattern	Without disabilities	With disabilities	Count	Proportion
111	5	0	5	.046
112	22	8	30	.275
121	2	0	2	.018
122	27	0	27	.248
211	1	0	1	.009
212	6	0	6	.055
221	2	0	2	.018
222	35	1	36	.330

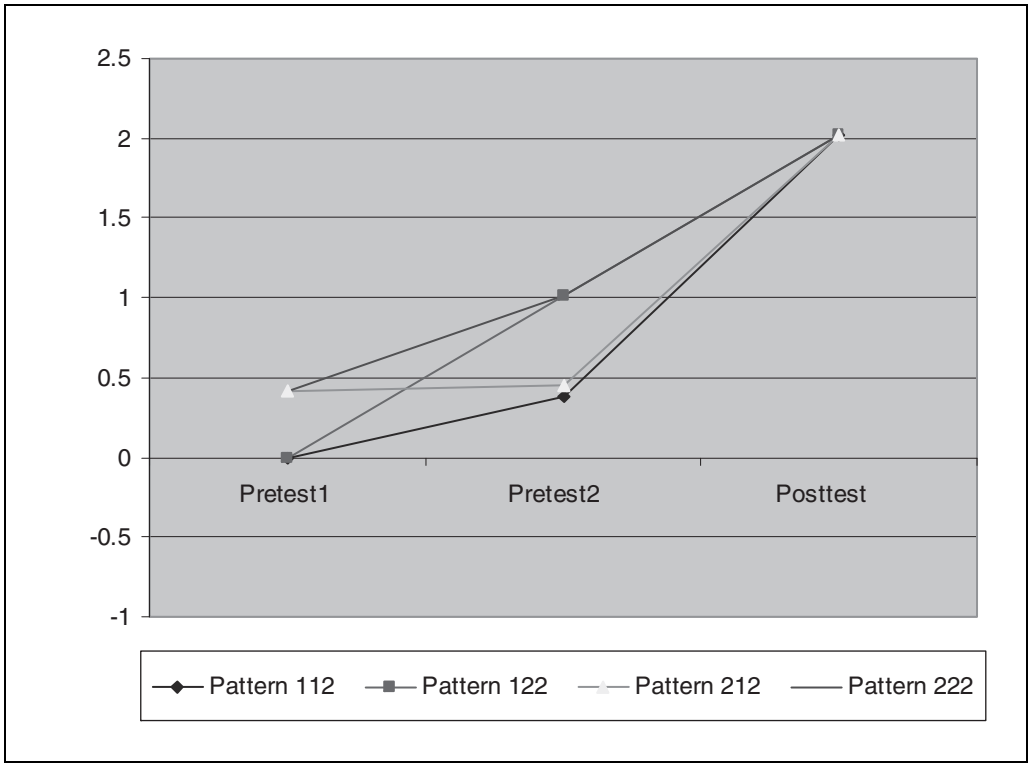

Figure 2. Ability mean profiles for the application.

there any change in ability over time (i.e., time effect)? The sample for the application was relatively small and, as a result, four of the eight transition patterns, 211, 221, 111, and 121, were excluded from this analysis because of very small numbers. A repeated measures analysis of variance was used to investigate the two hypotheses for patterns 112, 122, 212, and 222. An interaction was observed between time and group for transition patterns, $F(6, 190) = 5.65$, $p < .01$, partial eta square $= .151$, indicating that change in ability over time was different for different transition patterns. Because there was an interaction effect, the time effect was investigated for each transition pattern. For each pattern, a significant change in ability (p values <.01; partial eta square values ranged from .751 to .991) was observed across the three time points. Figure 2 shows the ability mean profile over three time points for patterns 112, 122, 212, and 222.

Table 9. Transition Probabilities for Each Latent Class for the Real Data Study

	Time 2	
Time 1	Latent Class	
Latent Class	1	2
1	.550	.450
2	.166	.834
	Time 3	
Time 2	Latent Class	
Latent Class	1	2
1	.174	.826
2	.085	.915

Latent transition probabilities. Estimated transition probabilities are shown in Table 9. Of the students who were in Class 1 at Pretest 1, 45% were subsequently classified into Class 2 at Pretest 2. This is an interesting result as no FOC intervention occurred until after Pretest 2. It is possible that this transition probability reflects a memory effect as the same performance tasks were used for all three time periods. Another possibility is that the Kim's Komet instructional intervention presented to these same students just prior to Pretest 2 may have had some effect. Second, 92% of the students who were in Class 1 at Pretest 2 were subsequently classified into Class 2 at posttest. Because FOC instruction was presented just before the posttest, the interpretation would appear to be that this transition probability reflects the impact of the FOC intervention.

Discussion

In this study, the LTA-MRM model was applied to a curriculum-based test of mathematics achievement to determine the effects of an instructional intervention. The LTA-MRM enables researchers to consider both the potential heterogeneity in response to intervention as well as a methodology for assessing the effects of the intervention over time.

The simulation study was done to examine the behavior of the LTA-MRM for fixed and joint estimation, use of a constraint of 0 for the covariance terms (implemented with the small sample and short test condition), test length, and sample size. Joint estimation yielded smaller RMSEs for item difficulty, group proportions, and ability than fixed estimation. Setting covariance terms to 0 for the small data set and short test condition also resulted in smaller RMSEs at each time point for group proportions, transition probabilities, and ability estimates than without the constraint. As the number of items and examinees increased, group proportions for each time point and RMSEs of transition probabilities decreased. Recovery of ability likewise improved with an increase in sample size.

In the simulation study, growth was simulated by increasing the proportions of high-ability group membership for each subsequent time point. Using this approach to generating data resulted in no differences between fixed and joint estimation in the detection of latent structure. It would be helpful to further compare the performance of fixed and joint estimation for the LTA-MRM by manipulating both change in terms of mean ability and change in terms of the proportions of latent class memberships. Simultaneously simulating both types of change should provide useful evidence about the two estimation methods.

Maximum likelihood estimation of class-specific parameters was difficult for the small sample size in the example, even though the LTA-MRM moderates the data sparseness problem better than the LTA-LCM. The LTA-MRM itself presents a computational problem, however, because it requires high dimensional integration as the number of time points increases. Joint

estimation for 14 items, 109 examinees, and three time points required 12 hours to complete a single simulation on a computer equipped with a 3.19-GHz processor with 3.00 GB of RAM. One possible approach to solving these problems could be using a Markov chain Monte Carlo (MCMC) algorithm to estimate model parameters. The use of mildly informative priors with MCMC on the probabilities of mixtures may help mitigate problems such as sparseness of data. Selection of appropriate priors on these probabilities can be potentially useful for bounding the mixture posterior away from nonidentifiability. Some problems remain, however, such as model identification and label switching for either MLE or MCMC.

The purpose of this article was to demonstrate the potential for the LTA-MixIRTM method for studying change in a behavioral process. This was done in the context of a problem that used the same set of performance tasks (i.e., items) over multiple time points. It is possible in such a case that memory may have played a part in test responses for some students. There was some evidence, in fact, that memory effects may have been present in one of the transition patterns. Memory effects, response consistency effects, and practice effects are all problems that may exist when dealing with repeated measures, possibly resulting in violation of the local independence assumption within the each pattern. In the application, however, the test items were performance tasks with multiple steps embedded in each one. For some of the items, students had to interpret schematic plans of a building project, figure out the most economical use of wood and other materials, and compute the total cost. Students were not told how they did on the pretests or posttests, nor were they shown the correct answers. Thus, it is unlikely that memory of test items had much, if any, impact on overall student test performance.

Extension of the LTA-MRM to modeling of different items across multiple time points is something that also might usefully be considered. This would require locating class-specific item parameter estimates on a common scale along with (some) anchor items. As was described for the example, class-specific item difficulties can be calibrated separately and then fixed as the first step in the estimation of an LTA-MRM. A joint estimation can then be used in the second step by setting class-invariant items as anchor items.

Results from the simulation study suggested that it was possible to obtain stable estimates for the LTA-MRM, given the sample size in the empirical data set. For larger data sets, it should be possible to incorporate more complex measurement models such as a mixture two-parameter or a mixture three-parameter model into an LTA.

In the example, it was found that items requiring two skills were clearly more difficult for members of Class 1 after class-specific item difficulty estimates were put onto the same metric. Although differential item functioning (DIF) in some items is indicated between the two latent classes, as Maij-de Meij, Kelderman, and van der Flier (2008) note, the presence of DIF does not necessarily imply that the constructs being measured in each latent class are necessarily different. However, it is still the case that ability estimates may reflect somewhat different response processes across classes. As Embretson and Reise (2000) note, it is still not clear whether abilities should be interpreted equivalently across classes, or whether it is necessary to adjust scores for deficient skill states.

Acknowledgments

The authors thank two anonymous reviewers and the editor for their insightful comments and suggestions.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: Partial funding by U.S. Department of Education, Institute of Education Sciences Cognition and Student Learning Research Program (R305H040032). Any opinions, findings, or conclusions are those of the authors and do not necessarily reflect the views of the supporting organizations.

References

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*, 3-16.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*, 381-409.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a Mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331-348.

Bottge, B. A., Heinrichs, M., Chan, S. Y., Mehta, Z. D., & Watson, E. (2003). Effects of video-based and applied problems on the procedural math skills of average-and low-achieving adolescents. *Journal of Special Education Technology, 18*, 5-22.

Bottge, B. A., Rueda, E., Serlin, R. C., Hung, Y., & Kwon, J. M. (2007). Shrinking achievement differences with anchored math problems: Challenges and possibilities. *Journal of Special Education, 41*, 31-49.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with applications to DIF. *Journal of Educational and Behavioral Statistics, 35*, 336-370.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2006, June). *An investigation of priors on the probabilities of mixtures in the mixture Rasch model*. Paper presented at the International Meeting of the Psychometric Society: The 71st annual meeting of the Psychometric Society, Montreal, Canada.

Clifford, M. M. (1991). Risk taking: Theoretical, empirical, and educational considerations. *Educational Psychologist, 26*, 263-297.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

Cohen, A. S., Cho, S.-J., & Kim, S.-H. (2005, April). *A mixture testlet model for educational tests*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Collins, L. M., & Flaherty, B. F. (2002). Latent class models for longitudinal data. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 287-303). Cambridge, UK: Cambridge University Press.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27*, 131-157.

Congdon, P. (2003). *Applied Bayesian modeling*. New York: John Wiley.

Darling-Hammond, L. (1996). The right to learn and the advancement of teaching: Research, policy, practice for democratic education. *Educational Researcher, 25*(5), 5-17.

De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review, 112*, 129-158.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495-515.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. N. M. de Gruijter & L. J. Th. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97-110). New York: John Wiley.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54*, 599-624.

Fischer, G. H. (1995). Linear logistic models for change. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 157-180). New York: Springer-Verlag.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.

Fuchs, L. S., & Fuchs, D. (2002). Mathematical problem-solving profiles of students with mathematics disabilities with and without comorbid reading difficulties. *Journal of Learning Disabilities, 35*, 563-573.

Geary, D. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin, 114*, 345-362.

Geary, D. C., Hamson, C. O., & Hoard, M. K. (2000). Numerical and arithmetical cognition: A longitudinal study of process and concept deficits in children with learning disability. *Journal of Experimental Psychology, 77*, 236-263.

Graham, J. W., Collins, L. M., Wugalter, S. E., Chung, N. K., & Hansen, W. B. (1991). Modeling transition in latent stage-sequential processes: A substance use prevention example. *Journal of Consulting and Clinical Psychology, 59*, 48-57.

Heinen, T. (1996). *Latent classes and discrete latent trait models*. Thousand Oaks, CA: Sage.

Knapp, M. S., & Turnbull, B. J. (1990). *Better schooling for the children of poverty: Alternatives to conventional wisdom* (No. 1). Washington, DC: U.S. Department of Education, Office of Planning, Budget and Evaluation.

Langeheine, R., & van de Pol, F. (2002). Latent Markov chains. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 304-341). Cambridge, UK: Cambridge University Press.

Lesh, R., & Kelly, A. (2000). Multitiered teaching experiments. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 197-230). Mahwah, NJ: Lawrence Erlbaum.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement, 33*, 353-373.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*, 611-631.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.

Muth, K. D. (1984). Solving arithmetic word problems: Role of reading and computational skills. *Journal of Educational Psychology, 76*, 205-210.

Muthén, B. O. (2008). Latent variable hybrids: Overview of old and new models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1-24). Charlotte, NC: Information Age.

Muthén, B., & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors, 31*, 1050-1066.

Muthén, B., Brown, C. H., Booil, J. K. M., Khoo, S.-T., Yang, C. C., Wang, C.-P., & Kellam, S. G. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics, 3*, 459-475.

Muthén, L. K., & Muthén, B. O. (1998-2006). *Mplus users guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2006). Mplus [Computer program]. Los Angeles, CA: Muthén & Muthén.

Parmar, R. S., Cawley, J. F., & Frazita, R. R. (1996). Word problem-solving by students with and without mild disabilities. *Exceptional Children, 62*, 415-429.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Rost, J. (1997). Logistic mixture models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449-463). New York: Springer.

University of Iowa. (2001). *The Iowa Tests of Basic Skills (ITBS, Form A)*. Ithaca, IL: Riverside.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont MA: Statistical Innovations.

von Davier, M. (2001). WINMIRA [Computer program]. St. Paul, MN: Assessment Systems Corporation.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*, 389-406.

Wilson, M. (1989). A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.

Yamamoto, K. Y., & Everson, H. T. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89-98). Münster, Germany: Wasmann.