PSYCHOMETRIKA—VOL. 78, NO. 3, 576–600
JULY 2013
DOI: 10.1007/S11336-012-9314-0

DETECTING INTERVENTION EFFECTS USING A MULTILEVEL LATENT TRANSITION ANALYSIS WITH A MIXTURE IRT MODEL

SUN-JOO CHO

VANDERBILT UNIVERSITY

ALLAN S. COHEN

THE UNIVERSITY OF GEORGIA

BRIAN BOTTGE

UNIVERSITY OF KENTUCKY

A multilevel latent transition analysis (LTA) with a mixture IRT measurement model (MixIRTM) is described for investigating the effectiveness of an intervention. The addition of a MixIRTM to the multilevel LTA permits consideration of both potential heterogeneity in students' response to instructional intervention as well as a methodology for assessing stage sequential change over time at both student and teacher levels. Results from an LTA–MixIRTM and multilevel LTA–MixIRTM were compared in the context of an educational intervention study. Both models were able to describe homogeneities in problem solving and transition patterns. However, ignoring a multilevel structure in LTA–MixIRTM led to different results in group membership assignment in empirical results. Results for the multilevel LTA–MixIRTM indicated that there were distinct individual differences in the different transition patterns. The students receiving the intervention treatment outscored their business as usual (i.e., control group) counterparts on the curriculum-based Fractions Computation test. In addition, 27.4 % of the students in the sample moved from the low ability student-level latent class to the high ability student-level latent class. Students were characterized differently depending on the teacher-level latent class.

Key words: latent class model, mixture item response theory model, multilevel latent transition analysis.

1. Introduction

Modeling of stage-sequential latent variables provides an alternative for analysis of some aspects of change. As an example, latent transition analysis (LTA, Collins & Wugalter, 1992) can be used to study an intervention in which the effect potentially differs within different latent classes in the data (Graham, Collins, Wugalter, Chung, & Hansen, 1991). The measurement model in LTA is the latent class model (LCM). LTA 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. LTA was extended to include a mixture item response theory model (MixIRTM) and was shown to be effective in detecting educational intervention effects in a simple instructional treatment (Cho, Cohen, Kim, & Bottge, 2010). The LTA with an item response theory (IRT) measurement model accounts for both a categorical latent variable (i.e., latent classes) and continuous latent variables (e.g., ability(-ies)). In this way, change can be observed as occurring both as a transition from one latent class to another, and as progress or decline along a latent continuous dimension within a latent class.

Requests for reprints should be sent to Sun-Joo Cho, Department of Psychology and Human Development, Peabody #H213A, Peabody College of Vanderbilt University, 230 Appleton Place, Nashville, TN 37203, USA. E-mail: sj.cho@vanderbilt.edu

The measurement model in the LTA was extended to a multilevel LTA (Asparouhov & Muthén, 2008; Yu, 2007) with an unrestricted LCM to account for dependency in item responses due to groups. Ignoring the multilevel data structure has been shown to lead to misclassification of latent group membership in LTA (Yu, 2007). As is the case with the LTA, the multilevel LTA model focuses on change that is homogeneous within latent class. Unrestricted LCMs 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), however, can be obtained by imposing constraints on LCMs. When the intent is to explore potentially useful hypotheses about structure in the data, it is generally more useful to use restricted LCMs. Furthermore, the MixIRTM often fits data better with a smaller number of latent classes than conventional LCM (Muthén & Asparouhov, 2006). In addition, as the number of latent classes increases, so does the number of transition patterns. For example, assume that LCM suggests a four-group solution and MixIRTM suggests a two-group solution as the best-fitting model at the student-level. With a two-group solution at the teacher level, there are 32 transition patterns in the multilevel LTA–LCM and 8 transition patterns in the multilevel LTA–MixIRTM with two time points. The large number of transition patterns makes it difficult to have enough observations in each transition pattern to estimate the model parameters unless the sample size is large. As test length increases, the matrix of response patterns also becomes increasingly sparse, making it more difficult to estimate model parameters. IRT model can potentially moderate some of this sparseness by applying a parametric model (i.e., a cumulative logistic or normal ogive function) with strong assumptions describing the relationship between the response probabilities and the continuous latent variable.

1.1. The Purpose of Study

In this paper, we use an extension of the multilevel LTA to include a MixIRTM. This extension has the potential to account for change in both categorical and continuous latent variables in the multilevel data. This is done by incorporating an IRT measurement model to a multilevel LTA, making it possible to address potential heterogeneity in change in categorical and continuous latent variables such as occurs in students' response to intervention at both individual and group levels. Asparouhov and Muthén (2008) introduced the two-level LTA by modeling the continuous latent variables at the group-level (e.g., teacher-level). In this study, we add the IRT measurement model to the two-level LTA. The resulting multilevel LTA–MixIRTM is then used to examine effects of a multi-group intervention study in which students are nested within teachers. A central question in this study is how to quantify an intervention effect.

We illustrate the use of the multilevel LTA–MixIRTM by answering the following substantive research questions in an educational intervention study: (i) Do students engage in problem solving differently? (ii) How do students' response processes change over time when, for example, they may differ with respect to different types of teachers? and (iii) How much progress do students make and does the progress depend on the intervention, on the teachers, or on the combination of teacher and intervention? Results from both the multilevel LTA–LCM and multilevel LTA–MixIRTM can be used to answer the first and second research questions. The multilevel LTA–LCM cannot be used to answer the third research question. In a multilevel LTA–MixIRTM, however, change can be accounted for as a transition from one latent class to another and as progress or lack thereof along a latent continuous variable (e.g., ability) within a latent class.

1.2. Mixture Models in an Intervention Study

The primary modeling framework of the multilevel LTA–MixIRTM is a mixture model. In this section, we describe how the mixture model can be used to detect an intervention effect when dealing with (i) evidence of multidimensionality in the post-test data, (ii) a multilevel data structure in which students are nested within teachers, and (iii) multiple groups such as control and treatment groups in an intervention study.

Multidimensional IRT models and MixIRTMs are two conceptually different approaches in IRT for modeling of multidimensionality in test data (Rijmen & De Boeck, 2005). Multidimensional IRT models are used to capture heterogeneity such as might arise due to item content, as when the various dimensions of a test are measured by different subgroups of items. MixIRTMs are intended to explain the heterogeneity in item responses among latent subgroups of examinees. MixIRTMs have been used to detect latent sub-populations that differ systematically in their responses on educational or psychological tests. For example, this methodology has been shown to be useful for identification of subgroups of examinees for which an item or a group of items function differently among groups as an indication of multidimensionality without the need of specifying these groups a priori (Rost, 1990). This approach models a continuous latent variable within latent subgroups of examinees. When there are more than two latent subgroups or latent classes in an intervention study, this can be thought of as reflecting differential effects of the intervention. The use of mixture modeling, however, does not by itself mean that there are distinct latent subgroups. It has been noted that mixture models (such as a mixture model portion of MixIRTM) can be used when there is multimodality in the latent variable (or factor) score distributions from multidimensional models (McDonald, 1967; Steinley & McDonald, 2007). McDonald (1967) recommended visually inspecting the latent variable score distributions to determine whether a continuous or categorical latent space is present. When multimodality exists in the latent variable score distributions, the use of mixture models can be supported.

When the intervention is a group randomized trial, multilevel models are needed to model the intervention effect, such as at student and teacher levels. Mixture models have been extended for data with a multilevel structure (Asparouhov & Muthén, 2008; Henry & Muthén, 2010; Vermunt, 2003, 2008). In the example illustrated in this paper, teachers were clustered into teacher-level latent classes with respect to different proportions of student-level latent classes.

As is shown in the example, manifest multiple groups, such as control and treatment groups, were not modeled; this differs from the case in which manifest multiple group models are used. Instead, the regression of the latent transition proportions on intervention status at the teacher level is modeled. The regression of the latent class variable on intervention status at the teacher level recognizes that teacher involvement in the intervention can affect the proportion of students in each class, thus answering the research question, "Does the intervention affect the probability that a student will be in a particular latent group as compared with a reference latent group?"

1.3. Related Models

Several recently developed longitudinal models contain features that are incorporated here into the multilevel LTA–MixIRTM. These include (i) modeling a continuous latent variable within a mixture, (ii) use of a latent Markov process for longitudinal data, and (iii) inclusion of a multilevel data structure. Cho et al. (2010) and Rijmen, De Boeck, & van der Maas (2005) included (i) modeling a continuous latent variable within a mixture and (ii) use of a latent Markov process for longitudinal data. von Davier, Xu, & Carstensen (2011) used (i) modeling a continuous latent variable within a mixture and (iii) inclusion of a multilevel data structure. Yu (2007) and Asparouhov and Muthén (2008) employed (ii) use of a latent Markov process for longitudinal data and (iii) inclusion of a multilevel data structure. However, no models have yet been developed to simultaneously account for all three features.

Below, a host study is described to show how the multilevel LTA–MixIRTM can be used. Next, a description is presented of the multilevel LTA–MixIRTM followed by a discussion of its estimation. Subsequently, the results of an empirical study are presented.

2. Effects of Enhanced Anchored Instruction

2.1. On Mathematics Achievement

A host study funded by the Institute of Education Sciences (U.S. Department of Education) was used to illustrate use of the multilevel LTA–MixIRTM. The purpose of the host study was to test the efficacy of *Enhanced Anchored Instruction* (EAI), which Bottge and his colleagues designed for improving the computation and problem solving skills of adolescents, including those of low-performing students with learning disabilities. The description below is intended to provide only an overview of the study; and, thus, it leaves out details included in the full report (Bottge, Ma, Toland, Gassaway, & Butler, 2012).

2.2. Study Design

The study employed a pre-test–post-test cluster-randomized school-based trial to compare the effects of two instructional conditions (EAI vs. Business As Usual [BAU]) on students' ability to compute and problem solve. EAI is modeled after anchored instruction (AI; Cognition and Technology Group at Vanderbilt, 1990, 1997; Hickey, Moore, & Pellegrino, 2001), an instructional method that typically includes short 8–15 minute video-based scenarios in which adolescents are portrayed attempting to solve an interesting problem. Teachers ask students probing questions and offer instructional guidance as they search scenes in the video to identify relevant information for helping the adolescents solve the problem. This way of portraying problem contexts does not require students to decode and comprehend text and thus eliminates the need for reading, a skill that many low-achieving math students also lack. EAI also includes more practical, hands-on applications to help students "visualize" the abstract concepts. The newest unit is a set of computer-based modules and manipulatives to boost students' understanding of and computation with fractions.

2.3. Teacher and Student Samples

Because schools scheduled students to classes, random assignment was by school rather than at the class or student level to avoid treatment contamination. A total of 49 special education teachers from 31 middle schools and their students with disabilities in math (MD) in Grades 6, 7, and 8 participated in the study. In all, 15 EAI schools (23 teachers, 33 resource rooms) and 16 BAU schools (26 teachers, 32 resource rooms) completed the study.

Students received their math instruction from special education teachers in self-contained special education settings because they lacked the prerequisite skills for learning grade-level content in general education math classes. Students received special education services for a disability in one of three categories: Mild Mental Disability (MMD), Other Health Impairment (OHI), or Specific Learning Disability (SLD). A smaller number of students were receiving services for Autism and Emotional Behavioral Disabilities (EBD). A total of 407 participated students in the study. Of these, 97 students did not respond to both the pre-test and post-test. As a result, 310 students remained in the final sample. These students were nested within the 49 teachers in the study. The smallest number of students analyzed for a teacher was one, and the largest was 21.

Chi-square tests of equal proportions indicated teachers were comparable across instructional conditions (BAU vs. EAI groups) in gender, ethnicity, and education level. Groups were also similar in years of teaching experience. Most teachers were white, female, well-educated, and experienced. Students were also comparable across instructional conditions in gender, ethnicity, subsidized lunch, and disability area.

Teachers assigned to the EAI condition taught five units that addressed several of the Common Core State Standards Initiative—Mathematics (CCSSI-M, 2011): Ratios and Proportional

Relationships, Number System, Statistics and Probability, and Geometry. They included Fractions at Work (FAW), a computer-based module and concretized manipulatives used to develop students' performance in computing fractions and four problem-solving units (i.e., two video-based anchored problems, and two related hands-on applied projects). The analysis in this paper deals primarily with the outcomes of FAW, which contained seven chapters designed to help students understand critical concepts (e.g., fraction equivalence) and to teach students how to compute with fractions (e.g., addition and subtraction). Teachers completed the FAW unit in an average of four weeks.

2.4. Measures: Fractions Computation Test

Two researcher-developed tests and two standardized achievement tests assessed the effects of EAI on the students' computation and problem-solving skills. In this paper, only the researcher-developed *Fraction Computation Test* was used to demonstrate the utility of the multilevel LTA–MixIRTM. The 20-item (14 addition, 6 subtraction) test measured students' ability to add and subtract fractions. Calculator use was not allowed. These 20 items have different item features. Items asked students to add and subtract fractions with *like* denominators (e.g., $\frac{1}{4}+\frac{3}{4}$), *unlike* denominators where the larger denominator could serve as the common denominator (e.g., $\frac{7}{8}-\frac{1}{4}$), and *unlike* denominators where neither could serve as the common denominator (e.g., $8\frac{2}{9}+2\frac{1}{2}$). Items also differed on whether or not the fractions could be found on the markings of a ruler (e.g., $\frac{3}{4}+\frac{7}{8}$) or not found on the ruler (e.g., $\frac{1}{5}+\frac{1}{3}$). All items except Item 1 could be found on the marking of a ruler. The test included *simple* fractions (e.g., $\frac{3}{8}+\frac{3}{4}$) and *mixed* numbers as well as addition of three fractions (e.g., $4\frac{1}{16}+\frac{1}{8}+\frac{1}{2}$). Items also differ on whether or not they contained two stacks (e.g., $\frac{\frac{3}{4}}{\frac{1}{2}}$) or three stacks (i.e., one more stack in the two-stack example).

There were a total of 42 points on the test. For 18 of the 20 items, students could earn 0, 1, or 2 points. On two items, students could earn 3 points, if they simplified the answer (i.e., revised the fraction to simple terms). Cronbach's alpha internal consistency estimates of previous versions of this test with the 42 points were 0.94 (Bottge, 1999), 0.91 (Bottge, Heinrichs, Mehta, & Hung, 2002), and 0.97 (Bottge, Heinrichs, Mehta, Rueda, Hung, & Danneker, 2004; Bottge, Rueda, Serlin, Hung, & Kwon, 2007). Internal consistency estimates for this sample were 0.81 at pre-test and 0.96 at post-test. Interrater agreement on a random 20 % of the tests was 97 %. Less than 1 % of students in the sample received partial scores (i.e., score 1 for 18 items and scores 1 or 2 for two of the items) on any of the items on the test. In this paper, binary responses were considered, 1 for correct responses and 0 for incorrect responses. Partial scores were also considered as correct responses. There was one missing item response from a single student in the final sample of 310 students. There was one missing item response from a single student in the final sample of 310 students.

On the pre-test, 51 % of students got scores of 0 in the BAU group and 60 % of students got scores of 0 in the EAI group. Means and standard deviations of the total scores (which ranged from 0 to 20 points) were 1.616 and 2.478 for the BAU group and 1.219 and 1.764 for the EAI group, respectively. These descriptive statistics show that most students had minimum scores on the pre-test. On the post-test, there was more variation in total scores in the EAI group than in the BAU group: 37 % of students received scores of 0 in the BAU group and 12 % received scores of 0 in the EAI group. Means and standard deviations of the total scores were 2.378 and 3.201 for the BAU group and 8.918 and 6.954 for the EAI group, respectively. The mean of the dependent variable differs significantly among the levels of program type.

Scorers also identified and coded the primary error students made on each incorrect item. Error codes were generated based on a procedure that included identifying errors on a sample of pretests, categorizing and labeling the codes, and scoring a second sample to assess each code's descriptive value. This resulted in a final total of 11 codes, as shown in Appendix.

3. Models

The models described below focus on binary responses.

3.1. Multilevel Latent Transition Model with a Mixture Item Response Model

In this section, we describe a MixIRTM and then show how it can be used in a multilevel LTA–MixIRTM.

3.1.1. Mixture Item Response Theory Model A 2-parameter logistic (2PL) version of a MixIRTM (Rost, 1990) is described as follows: The probability of obtaining response pattern $\mathbf{y}_j$ for person j is

$$\begin{aligned} P(\mathbf{y}_j) &= \sum_{g=1}^{G} P(C_j = g) \cdot \prod_{i=1}^{I} P(y_{ji} = 1|g, \theta_{jg}) \\ &= \sum_{g=1}^{G} P(C_j = g) \cdot \prod_{i=1}^{I} \frac{\exp(a_{ig} \cdot \theta_{jg} - b_{ig})}{1 + \exp(a_{ig} \cdot \theta_{jg} - b_{ig})}, \end{aligned} \tag{1}$$

where j is a person index ($j = 1, \ldots, J$), i is an item index ($i = 1, \ldots, I$), C_j is a latent class variable denoting latent class membership, g is a specific individual-level categorical latent variable or latent class ($g = 1, \ldots, G$), $P(C_j = g)$ is the proportion of the population in latent class g, θ_{jg} is a continuous latent variable, a_{ig} is a class-specific item discrimination, and b_{ig} is a class-specific item intercept.

3.1.2. Multilevel Latent Transition Model with a Mixture Item Response Theory Model The probability of obtaining response pattern $\mathbf{y}_{js}$ for person j nested within group s across time points with a MixIRTM is given as

$$\begin{aligned} P(\mathbf{y}_{js}) = \sum_{h=1}^{H} P(D_s = h) \sum_{g_1=1}^{G_1} \cdots \sum_{g_T=1}^{G_T} P(C_{js,t=1} = g_1|D_s = h) \\ \cdot \prod_{t=2}^{T} P(C_{jst} = g_t|C_{js,t-1} = g_{t-1}, D_s = h) \left[\prod_{i=1}^{I} \prod_{t=1}^{T} P(y_{jsit} = 1|g_t, h, \theta_{g_t h}) \right], \end{aligned} \tag{2}$$

where t is a time point index ($t = 1, \ldots, T$), C_{jst} is a latent variable denoting individual-level latent class membership at a time point t, g_t is a specific individual-level latent class at a time point t ($g_t = 1, \ldots, G_t$), D_s is a latent variable denoting group-level latent class membership, h is a specific group-level latent class ($h = 1, \ldots, H$), $P(D_s = h)$ is the proportion of the population in group-level latent class h, $P(C_{js,t=1} = g_1|D_s = h)$ is the proportion of the population in latent class g_1 at Time 1, $P(C_{jst} = g_t|C_{js,t-1} = g_{t-1}, D_s = h)$ is a transition proportion, and $P(y_{jsit} = 1|g_t, h, \theta_{g_t h})$[1] is a measurement model. To identify the model, $\sum_{h=1}^{H} P(D_s = h) = 1$ and $\sum_{g=1}^{G} P(C_{jst} = g|D_s = h) = 1$. Constraints for distributions of $\theta_{g_t h}$ are explained in a later section (see also Table 1). Each of the terms of the multilevel LTA–MixIRTM are described below.

[1]To present the individual differences for persons j nested within group s at a time point t, $\theta_{g_t h}$ can be presented as $\theta_{jstg_t h}$. For simplicity, the subscripts, j, s, and t, are dropped.

TABLE 1.
Constraint labels on continuous latent variable structure for each transition pattern.

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

3.1.3. Latent Class and Transition Proportions $P(D_s = h)$ is the proportion of the population in a group-level latent class h, and is defined as follows:

$$P(D_s = h) = \frac{\exp(\gamma_h)}{\sum_{h=1}^{H} \exp(\gamma_h)}, \tag{3}$$

where γ_h is the log odds when comparing a group-level latent class h to a reference group-level class (in this study, $h = 1$). For identifiability, γ_h for the first group-level latent class was set to be 0.

The proportion of the population in latent class g_1 at Time 1, $P(C_{js,t=1} = g_1 | D_s = h)$, is defined as follows:

$$P(C_{js,t=1} = g_1 | D_s = h) = \frac{\exp(\gamma_{g_1 h})}{\sum_{g_1=1}^{G} \exp(\gamma_{g_1 h})}, \tag{4}$$

where $\gamma_{g_1 h}$ is the log odds when comparing an individual-level latent class g_1 to a reference individual-level latent class (in this study, $g_1 = 1$). For identifiability, $\gamma_{g_1 h}$ for the first individual-level class is set to be 0.

For $t = 2, \ldots, T$, the transition proportions are defined as

$$P(C_{jst} = g_t | C_{js,t-1} = g_{t-1}, D_s = h) = \frac{\exp(\gamma_{g_t h})}{\sum_{g_t=1}^{G} \exp(\gamma_{g_t h})}, \tag{5}$$

where $\gamma_{g_t h}$ is log odds when comparing an individual-level latent class g_t to a reference individual-level latent class ($g_t = 1$, in this study). For identifiability, $\gamma_{g_t h}$ for the first individual-level class is set to be 0.

In addition, a group-level covariate, x_s, indicating intervention condition, is employed to help explain transition proportions, $\gamma_{g_t h}$. The $\gamma_{g_1 h}$ and $\gamma_{g_t h}$ can be described as

$$\gamma_{g_1 h} = \beta_{10} + \beta_{1h}, \tag{6}$$

where β_{10} is an effect of average log-odds when comparing individual-level latent class g_1 to a reference individual-level latent class ($g_1 = 1$, in this study) at a time point 1 and β_{1h} is an effect of $(D_s = h)$:

$$\gamma_{g_t h} = \beta_{t0} + \beta_{t2h} + \beta_{t3} \cdot \gamma_{g_{t-1}} + \beta_{t4} \cdot x_s, \tag{7}$$

where β_{t0} is an effect of average log-odds when comparing individual-level latent class g_1 to a reference individual-level latent class ($g_1 = 1$, in this study), β_{t2h} is an effect of log-odds for a previous time point $t-1$ at a group-level latent class h at a time point t, β_{t3} is an effect of log-odds for a previous time point $t-1$ at an individual level at a time point t, and β_{t4} is an intervention effect at time point t.

3.1.4. Measurement Model Item parameter invariance across time points and latent classes at the group level was assumed to ensure that the number and structure of the latent classes were the same across time. Item parameter invariance across time points for all items is necessary to interpret the transition probabilities. Item parameter invariance across group-level latent classes for all items is necessary to make comparisons across group-level latent classes with respect to the different proportions of student-level latent classes. With the two assumptions on measurement invariance across time points ts and group-level latent classes hs, the measurement model in Equation (2) is

$$P(y_{jsit} = 1 | g_t, h, \theta_{g_t h}) = \frac{\exp(a_{ig} \cdot \theta_{g_t h} - b_{ig})}{1 + \exp(a_{ig} \cdot \theta_{g_t h} - b_{ig})}, \tag{8}$$

where a_{ig} and b_{ig} are a time invariant and class-specific item discrimination and intercept respectively. Multidimensional ability ($[\theta_{g_1 h}, \ldots, \theta_{g_T h}]'$) modeling over time in the multilevel LTA–MixIRTM is the same as in Anderson (1985), as within each pattern abilities are time-specific at the individual level. $\theta_{g_1 h}$ can be considered as the initial ability. $\theta_{g_t h}$ for $t = 2, \ldots, T$ involves initial ability and changes at each time, t. Let the ability at each time point be $\theta_{g_t h}$ for each pattern and the change in ability at time t be $\theta^{(+)}_{g_t h}$. For the first time point, $\theta_{g_1 h} = \theta_{g_t h}$. For $t = 2, \ldots, T$, $\theta_{g_t h} = \theta_{g_{t-1} h} + \theta^{(+)}_{g_t h}$ holds. The measurement model is a multidimensional model at a given time point at which item parameters, represented by a_{ig} and b_{ig}, may differ across individual-level latent classes gs.

The multidimensional ability structure in a multilevel LTA–MixIRTM is given as

$$[\theta_{g_1 h}, \ldots, \theta_{g_T h}]' \sim MN(\vec{\mu}_{g_t h}, \Sigma_{g_t h}), \tag{9}$$

where $\vec{\mu}_{g_t h}$ is the $T \times 1$ mean vector and $\Sigma_{g_t h}$ is the $T \times T$ variance–covariance of the ability dimension across time points in a multivariate normal (MN) distribution for a particular pattern of latent classes. In this study, an unrestricted covariance structure for $\Sigma_{g_t h}$ is modeled.

3.1.5. Estimation of Latent Structure and Scale Comparability Item response probabilities in the latent variable structure described above have a continuous latent variable with a multidimensional structure over time points for each transition pattern. Thus, the probability of a correct response, $P(y_{jsit} = 1 | g_t, h, \theta_{g_t h})$ can be different across persons and time points, when the same items are administered across time points.

In this study, class-specific item parameters were estimated using post-test data. This was because most students were only able to solve a very small subset of the items on the pre-test, namely those items with *like* denominators. This use of class-specific item parameters 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 parameter estimates were then held constant across time points.

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 MixIRTM, class-specific item parameters as well as group memberships g and h are estimated simultaneously.

4. Estimation

The computer program Mplus (Muthén & Muthén, 2006–2010) was used to estimate parameters of all models in this study. Mplus input files are available from the first author upon request. TYPE = TWOLEVEL MIXTURE causes Mplus to estimate parameters of a multilevel mixture model. Maximum likelihood estimation is implemented in Mplus (ESTIMATOR = ML) with the numerical integration option (ALGORITHM = INTEGRATION). Maximum likelihood optimization was done in two stages. First, in this study, an optimization was carried out for 100 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 = TWOLEVEL MIXTURE.

The primary modeling framework of the multilevel LTA–MixMRM 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, we discuss some of these issues relative to estimation of the multilevel LTA–MixIRTM.

Non-identification of a model implies that different parameter estimates yield the same log-likelihood value. Since the likelihood function is invariant with respect to different permutations of model parameters in mixture modeling, there is a problem that can arise, called label switching. Two types of label switching occur with mixture modeling (Cho, Cohen, & Kim, 2012b). The first type occurs across iterations within a single chain in a Bayesian solution. The second type of label switching can happen for both Bayesian and maximum likelihood estimation. The second type does not distort the final estimates using a single data set in an empirical study, but one needs to be aware of it, when interpreting results of simulation studies and interpreting results of empirical studies with alternative models.

Constraints on parameters, such as ordering restrictions on mixing proportions (McLachlan & Peel, 2000), can be imposed to prevent label switching. In our applications using maximum likelihood estimation, the possibility of label switching was examined after parameter estimates were obtained with constraints on population parameters of the multivariate ability structure (these constraints are described below, in the section Constraints on Parameters in LTA–MixIRTM and Multilevel LTA–MixIRTM). When label switching occurs, labels can be renamed by matching the patterns of estimates across alternative models or replications. For example, assume the generated individual-level mixture proportions were 0.7 and 0.3 in a two-group model. From one simulated data set based on this model, the estimated mixture proportions might be 0.65 and 0.35 for Classes 1 and 2, respectively. From a second simulated data set based on this model, they might be 0.25 and 0.75. Looking at the latter case, one could conclude that the label switching occurred. Therefore, prior to investigating the recovery of parameters from these two simulated data sets, labels of classes from the second simulated data set should be renamed as Classes 2 and 1, respectively. Post-hoc inspection for label switching can sometimes be problematic, however, when the true parameters are similar across latent classes (e.g., if mixture proportions were 0.5 for both classes in a two-group model). In this study, the patterns of estimates across latent classes for all parameters in the models were inspected. In addition, standard

deviations of class-specific parameter estimates across alternative models were investigated as an additional indication of the possibility label switching had occurred.

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 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 (McLachlan & Peel, 2000; Muthén, Brown, Jo, Khoo, Yang, Wang, Kellam, Carlin, & Liao, 2002; Frühwirth-Schnatter, 2006).

The usual method used for checking whether the model is identified or a local solution has been obtained is to run the model with multiple different starting values (McLachlan & Peel, 2000). Observing the same log-likelihood from multiple set of initial values increases confidence that the solution is not local. In this study, 100 different sets of initial values were set to check model identification and local maxima.

To specify the multidimensional (continuous) latent variable structure, means and variances of the multidimensional distribution were set to be equal and individual-level membership was assumed to be the same at each time point within a group-level latent class. Covariances also were constrained to be equal as an individual-level group membership was assumed to be the same at adjacent time periods within a group-level latent class. Table 1 shows the labels of constraints on the latent variable structure for each transition pattern for the two time periods in this study: two latent classes at the individual level ($G = 2$) and two latent classes at the group level ($H = 2$). Transition Pattern 111 indicates group-level Class 1, and individual-level Class 1 at Time 1 and individual-level Class 1 at Time 2. The values in Table 1 are labels indicating which of the different terms in the model were constrained to be equal. There are 8 unique labels, for example, for the ability means (i.e., 1, 2, 3, 4, 5, 6, 7, and 8). These labels indicate that the means with the same label were constrained to be equal. So, for transition Patterns 111 and 112, the label 1 indicates that means at Time 1 for these two patterns were constrained to be equal. Similarly, for group-level Class 1, there are 4 labels in this table for ability variances (i.e., 9, 10, 11, and 12) and covariances (i.e., 13, 14, 15, and 16) indicating which variances and covariances were constrained to be equal. For group-level Class 2, there are 4 labels for ability variances (i.e., 17, 18, 19, and 20) and covariances (i.e., 21, 22, 23, and 24) indicating which variances and covariances were constrained to be equal. As model identification constraints, means and variances of the multidimensional latent variable structure for the Pattern 111, are set to be 0s and 1s, respectively. These labels are shown in Figure 1.

5. Empirical Study Results

In this section, we show how results of a multilevel LTA–MixIRTM can be used to help explain the response to intervention in a group-randomized trials design. Results for the LTA–MixIRTM (Cho et al., 2010; Rijmen et al., 2005) and the multilevel LTA–MixIRTM were compared to show how the student-level latent class assignment differs when the teacher-level differences are ignored.

5.1. Multilevel Data Structure

Intraclass correlations (ICCs, Raudenbush & Bryk, 2002) were used to investigate the multilevel structure of the data. The ICC based on results of the multilevel 1-parameter logistic model (Cho & Rabe-Hesketh, 2011) were 0.117 and 0.406 for pre-test and post-test, respectively, indicating 11.7 % and 40.6 % of the total variance in ability was explained at the teacher level for pre-test and post-test, respectively. The increase in variance on the post-test occurred for teachers in both latent classes mainly from the EAI intervention effect and to a lesser extent from instruction in the BAU condition.

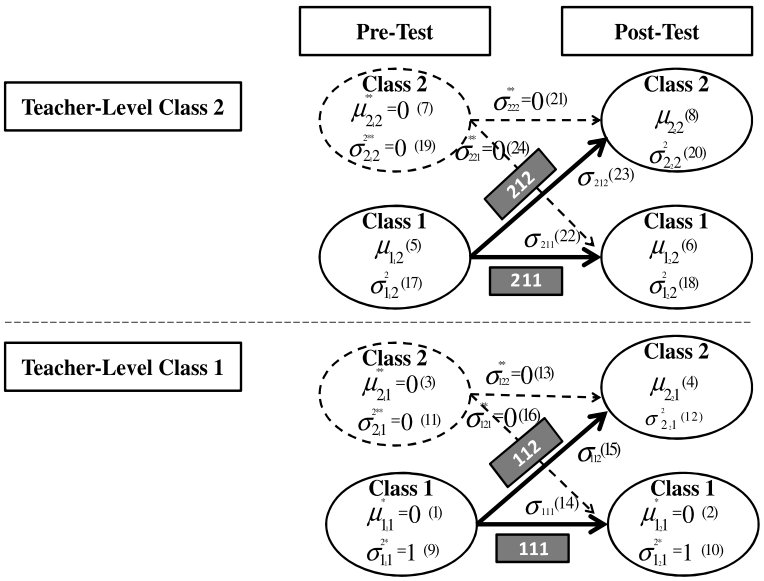

FIGURE 1.
Multidimensional ability structure of a multilevel LTA–MixIRTM with constraints.

5.2. *Detection of Latent Classes*

We first determined the number of latent classes at the student level and then selected the number of latent classes at the teacher level (see Henry & Muthén, 2010; Lee Van Horn, Fagan, Jaki, Brown, Hawkins, Arthur, Abbott, & Catalano, 2008). As noted above, analysis of the pre-test data indicated that most students were only able to solve items with *like* denominators. This is understandable because students with MD do not often receive much instruction beyond whole number computation. Students were more likely to solve more items correctly after intervention. Thus, it might be expected there would be more variability in item response patterns after than before the intervention. Analysis of the post-test data, in fact, indicated that this occurred. Thus, post-test data were used to investigate the number of latent classes.

Results of a multilevel 2-dimensional 2PL IRT model (with two student-level dimensions and two teacher-level dimensions) indicated multimodality was present in ability score distributions at both the student and teacher levels, thus supporting the mixture modeling in our application (McDonald, 1967; Steinley & McDonald, 2007).

Table 2 shows the model selection results for the LCM and MixIRTM for several information criteria including the Bayesian information criterion (BIC: Schwarz, 1978), Akaike's information criterion (AIC: Akaike, 1973), Akaike's information criterion 3 for small samples (AIC3: Sugiura, 1978), and the consistent Akaike information criterion (CAIC: Bozdogan, 1987). Previous research has suggested that AIC tended to overestimate the number of latent classes in LCM (Nylund, Asparouhov, & Muthén, 2007; Yang, 2006) and in MixIRTM (Li, Cohen, Kim, & Cho, 2009; Nylund et al., 2007). CAIC has been found to perform better than AIC, but has been shown to be sensitive to the combination of unequal sample sizes and small sample sizes in LCM (Nylund et al., 2007). BIC has been show to perform the best among information criteria for selecting the model with the "correct" number of latent classes in both LCM (Nylund et al., 2007) and MixIRTM (Cho et al., 2012b; Li et al., 2009). Given these results, BIC was chosen as the primary information criterion for use in selecting the model at the student level. As shown in Table 2 (upper), AIC selected a model with one more latent classes than BIC for both LCM and MixIRTM. There were no consistent patterns, however, in model selection for either LCM

TABLE 2.
Model selection results.

Selection of student-level model

Model	Num. of LCs	Num. of parameters	Log-L	Model fit			
				BIC	AIC	AIC3	CAIC
LCM	1	20	−3201.546	6517.693	6443.091	6463.091	6463.091
	2	41	−1804.245	3843.424	3690.490	3731.490	3731.490
	3	62	−1620.160	3595.587	3364.321	3426.321	3426.321
	4	83	−1529.918	**3535.435**	3225.836	3308.836	3308.836
	5	104	−1492.093	3580.116	3192.185	3296.185	3296.185
MixIRTM	1	40	−1576.948	3383.100	3233.896	3273.896	3423.100
	2	81	−1462.138	3388.413	3086.275	3167.275	3469.413
	2*	79	−1462.519	**3377.717**	3083.039	3162.039	3456.717
	3	122	−1397.755	3494.581	3039.509	3161.509	3616.581
	4	163	−1372.327	3678.661	3070.655	3233.655	3841.661

*Item 15 as an anchor item.

Selection of teacher-level model for 2-group student-level model

Model	Num. of LCs	Num. of parameters	Log-L	Model fit			
				BIC	AIC	AIC3	CAIC
Multilevel	1	79	−1462.519	3377.717	3083.039	3162.039	3456.717
MixIRTM	2	81	−1444.711	**3353.560**	3051.421	3132.421	3434.560
	3	83	−1442.160	3359.919	3050.321	3133.321	3442.919

or MixIRTM with either AIC3 or CAIC. Using BIC as the criterion, the LCM analysis suggested a four-group solution. An analysis for the MixIRTM using BIC as the criterion suggested a two-group solution.

Class-invariant constraints on item parameters for scale comparability and model identification constraints were required in order to compare parameter estimates from the different latent classes in the MixIRTM. 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 significant change in the −2 log-likelihoods was observed between the all-items-fully-relaxed model and the one-item-constrained model for Item 15. This was taken to indicate that Item 15 could be used as an anchor item. With Item 15 as the anchor item, BIC still suggested the two-group solution.

Given the two-group student-level solution for the MixIRTM, a multilevel MixIRTM was fit to the post-test data. Results for BIC, AIC, AIC3, and CAIC are reported in Table 2 (bottom). Application of information criteria requires calculation of a penalty term such as the number of persons or the number of groups or both. Research has suggested that it is difficult to define the number of observations used for calculation of BIC in multilevel models (Skrondal & Rabe-Hesketh, 2004, p. 265). In multilevel item response models, including their extensions, in other words, it is not clear what quantities (i.e., number of persons, number of groups, or both) should be used to determine the penalty term used in BIC. When the number of persons has been used for this purpose in multilevel MixIRTM applications, BIC has been found to perform better than AIC. In addition, AIC has been shown to overestimate the number of latent classes for the multilevel MixIRTM (Cho & Cohen, 2010). BIC was chosen for this study, therefore, for selecting the number of latent classes at the teacher-level. An exploratory analysis suggested

the two-group teacher-level solution was the better fit based on BIC, AIC3, and CAIC. A three-group teacher-level solution was recommended by AIC (see Table 2). This result indicated that the proportions of student-level latent classes differed within teacher-level class. Using BIC as the criterion, teacher-level Class 1 had more students from student-level Class 1, and teacher-level Class 2 had more students from student-level Class 2.

5.3. Constraints on Parameters in LTA–MixIRTM and Multilevel LTA–MixIRTM

Figure 1 shows the different constraints on multidimensional ability structure for the multilevel LTA–MixIRTM at the student-level and teacher-level. Different constraints on parameters were considered to identify the model (indicated as $*$) and to take into account the data sparseness in the pre-test (indicated as $**$).

In Figure 1, Pattern 111 indicates teacher-level Class 1, student-level Class 1 at Time 1, and student-level Class 1 at Time 2. To identify the model, constraints were set on means ($\mu^{*}_{1_1 1} = 0$, $\mu^{*}_{1_2 1} = 0$) and variances ($\sigma^{2*}_{1_1 1} = 1$, $\sigma^{2*}_{1_2 1} = 1$) of the multidimensional ability distribution for Pattern 111.

From the descriptive analysis of the data, it was evident that most students were only able to solve items with *like* denominators (i.e., easy items) on the pre-test. In addition, a model with only one class at the student-level was found to fit the data based on BIC. Since it was expected that very few students, if any, would be detected in student-level Class 2 (the high-ability group) on the pre-test, the means ($\mu^{**}_{2_1 1} = 0$, $\mu^{**}_{2_1 2} = 0$) and variances ($\sigma^{2**}_{2_1 1} = 0$, $\sigma^{2**}_{2_1 2} = 0$) of ability for student-level Class 2 at Time 1 for both teacher-level latent classes were set to 0s (indicating that everyone in student-level Class 2 at Time 1 had a score of 0). This was used as a scale anchor point. In addition, covariances of the multidimensional ability distribution for Patterns 121, 122, 221, and 222 were set to 0s. With these constraints, four transition Patterns, 111, 112, 211, and 212, could potentially be observed in the multilevel LTA–MixIRTM. The same constraints, albeit without the multilevel structure, were used for the LTA–MixIRTM.

Labels of constraints illustrated in Table 1 are included in Figure 1 to show which labels are equal because group membership was assumed to be the same at each time point. In Figure 1, for example, Patterns 111 and 112 share the same student-level Class 1 in teacher-level Class 1 at the pre-test. Thus, these two patterns share the constraints for model identification, $\mu^{*}_{1_1 1} = 0(1)$ and $\sigma^{2*}_{1_1 1} = 1(9)$ in the figure. As another example, Patterns 221 and 222 share the same student-level Class 2 in teacher-level Class 2 at the pre-test. Consequently, these two patterns share the constraints for model structure, $\mu^{**}_{2_1 2} = 0(7)$ and $\sigma^{2**}_{2_1 2} = 0(19)$ in the figure.

5.4. Comparisons Between Multilevel LTA–MixIRTM and LTA–MixIRTM

The transition patterns and their multidimensional ability structures were compared between LTA–MixIRTM and multilevel LTA–MixIRTM, using the same item parameter estimates estimated with the MixIRTM, to show the effect of ignoring the multilevel structure. With the same item parameter estimates, labels of student-level latent classes also were the same between the models. At the pre-test, there was no difference in latent class assignment between multilevel LTA–MixIRTM and LTA–MixIRTM. A possible interpretation of this result is that students did not differ in their performance before intervention. The different classifications of students on the post-test, however, reflect the fact that students differed in their response to the intervention. The number of transition patterns and the ability distributions differed in this regard for the multilevel LTA–MixIRTM and the LTA–MixIRTM. Students were clustered into four transition patterns in the multilevel LTA–MixIRTM but were clustered into two transition patterns in the LTA–MixIRTM.

TABLE 3.
Counts and proportions of latent classes and transition patterns.

Teacher-level and student-level class counts

Latent class variable		Class	Count	Proportion
Teacher level		1	41	0.837
		2	8	0.163
Student level	Time 1	1	310	1.000
		2	0	0.000
	Time 2	1	225	0.726
		2	85	0.274

Transition patterns

Teacher class	Student class		Transition pattern	BAU frequency	EAI frequency	Total frequency	Total proportion
	Time 1	Time 2					
1	1	1	111	148	60	208	0.671
1	1	2	112	16	23	39	0.126
1	2	1	121	0	0	0	0.000
1	2	2	122	0	0	0	0.000
2	1	1	211	0	17	17	0.055
2	1	2	212	0	46	46	0.148
2	2	1	221	0	0	0	0.000
2	2	2	222	0	0	0	0.000

5.5. *Interpretations of Latent Classes and Intervention Effects with Results of Multilevel LTA–MixIRTM*

Table 3 shows class counts and proportions obtained by the estimated mixture proportions (Equations (3)–(7)). The models with a covariate, indicating the intervention grouping variable, and without a covariate on the transition proportions yielded the same group membership assignments at both student and teacher levels. The effect modeled by the intervention grouping variable on transition proportions, however, was significant ($\widehat{\beta}_{t4} = -3.778$, $SE_{\beta_{t4}} = 1.237$, $p \leq 0.002$). This indicated that the intervention affected the probability that a student was in student-level Class 2 rather than student-level Class 1 and that teacher involvement in the intervention affected the proportion of students in each class. A large estimate, $\widehat{\beta}_{t4} = -3.778$, was observed in Pattern 212; all 46 students in this pattern (i.e., in student-level Class 2 on the post-test) were from the EAI sample.

(i) *Do students engage in problem solving differently*? Given the data structure and model selection results, item parameters were estimated by fitting a two-group MixIRTM to the post-test data.

Item parameter estimates of the MixIRTM were fixed across the two time points. Table 4 shows the different item profiles across latent classes. Student-level latent classes were characterized by different item parameter profiles across latent classes. There was no overlap in item response curves between latent classes. The four *like* items were easier in student-level Class 1 than in student-level Class 2 (e.g., Item 17), while other items except these four items were easier in Class 2 than in Class 1 (e.g., Item 19). These results indicate that there was a qualitative difference between latent classes (De Boeck, Wilson, & Acton, 2005; Rost, 1990). The four *like* items were considered as easy items that students would be able to solve without the EAI intervention. More complex computation skills were required by the other items in order to handle the different kinds of numerical operations including the different types of denominators, types of

TABLE 4.
Item attributes and item parameter estimates (standard errors) for each student-level latent class.

Item	Attributes				Class 1		Class 2	
	Operation	Denominator	Type	Stacks	*a*	*b*	*a*	*b*
1	addition	like	simple	2	3.378(0.464)	−2.880(0.399)	1.275(0.361)	−0.371(0.490)
2	addition	like	simple	2	3.243(0.436)	−2.645(0.374)	1.419(0.406)	−0.374(0.466)
3	addition	unlike	simple	2	2.799(0.361)	1.481(0.275)	1.037(0.339)	−1.070(0.520)
4	addition	unlike	simple	2	2.947(0.385)	1.614(0.291)	1.126(0.315)	−1.506(0.536)
5	addition	unlike	simple	2	2.978(0.377)	2.201(0.325)	1.178(0.375)	−1.464(0.516)
6	addition	unlike	simple	2	3.016(0.381)	2.520(0.360)	1.249(0.394)	−0.770(0.431)
7	addition	unlike	mixed	2	2.952(0.373)	2.558(0.358)	1.112(0.355)	−0.917(0.481)
8	addition	unlike	mixed	2	3.012(0.390)	2.787(0.392)	1.093(0.351)	−1.288(0.472)
9	addition	unlike	mixed	2	2.756(0.348)	2.694(0.355)	1.089(0.366)	−0.712(0.460)
10	addition	unlike	mixed	2	3.661(0.454)	2.766(0.407)	1.977(0.540)	−0.160(0.331)
11	addition	unlike	simple	3	2.816(0.359)	2.426(0.343)	0.904(0.300)	−1.305(0.519)
12	addition	unlike	simple	3	2.663(0.340)	2.627(0.350)	1.139(0.387)	−0.371(0.412)
13	addition	unlike	mixed	3	3.016(0.374)	2.836(0.375)	2.009(0.568)	1.043(0.245)
14	addition	unlike	mixed	3	3.188(0.408)	3.135(0.415)	1.907(0.529)	−0.328(0.309)
15*	subtraction	like	simple	2	2.693(0.393)	−1.668(0.306)	2.693(0.393)	−1.668(0.306)
16	subtraction	unlike	simple	2	2.858(0.364)	2.195(0.323)	2.094(0.573)	0.411(0.271)
17	subtraction	like	mixed	2	2.202(0.331)	−0.451(0.229)	2.164(0.575)	0.189(0.293)
18	subtraction	unlike	mixed	2	1.190(0.199)	2.423(0.260)	1.229(0.439)	0.522(0.305)
19	subtraction	unlike	mixed	2	2.451(0.316)	2.643(0.346)	1.709(0.506)	0.434(0.256)
20	subtraction	unlike	mixed	2	1.933(0.274)	3.051(0.358)	1.400(0.484)	0.963(0.269)

*Item 15 as an anchor item.

computation, and numbers of stacks (i.e., two or three stacks). These skills were part of the focus of the EAI. Thus, student-level Class 1 was labeled as the low ability group and student-level Class 2 as the high ability group.

Characteristics of students classified into student-level Class 1 and student-level Class 2 based on the post-test results were examined to try to determine why students may have answered incorrectly. Table 5 provides proportions of correct answers and types of errors observed for each item in the post-test data across student-level latent classes for the "No Error" column and within a student-level latent class for the "Errors" columns. Types of errors and their examples are described in Appendix. The multilevel LTA–MixIRTM analysis revealed different patterns of errors for students in student-level Classes 1 and 2. For Items 1 and 2, the most common single error made by students in both student-level classes among errors was "Combining" (adding denominators and adding numerators). The next 12 items required students to add two *simple* fractions with *unlike* denominators (Items 3–6), add two *mixed* numbers with *unlike* denominators (Items 7–10), add three *simple* fractions with *unlike* denominators (Items 11–12), and add three *mixed* numbers with *unlike* denominators (Items 13–14). About half of the students in student-level Class 2 computed these problems correctly. The most common error made by students in both student-level latent classes on most items was not finding a common denominator, which reveals students' lack of basic misunderstanding of fractions. Even when students tried to find a common denominator, their work showed that they had difficulty computing it. The same pattern of errors emerged in subtracting fractions (Items 15–20).

Students in student-level Class 1, however, made more errors than students in student-level Class 2 for all items except for the "Equivalent Fraction Error" in Item 2 and the "Computation Error" in Items 10, 12, and 13. Item scorers were instructed to identify in each item the most important error or, in other words, the error that reflected a student's fundamental conceptual misunderstanding. Thus, if the mistake was identified as "Combining" or "Add All", the scorer did not look for any other errors. Because such a large proportion of students in student-level Class 1 made either "Combining" or "Add All" errors, there was much less chance of the scorer identifying a mistake as "Computation" or "Equivalent Fractions." Conversely, because fewer students in student-level Class 2 made "Combining" and "Add All" errors, there was a much greater chance of the scorer identifying one of the other errors.

(ii) *How do students' response processes change over time when, for example, they may differ with respect to different types of teachers*? As shown in Table 3 (top), all students were classified into student-level Class 1 (the low ability group) on the pre-test. On the post-test, 72.6 % were classified into student-level Class 1 and 27.4 % were classified into student-level Class 2 (the high ability group). Students who transitioned from student-level Class 1 on the pre-test to student-level Class 2 on the post-test were those who were able to solve items requiring the more complex computation skills that were emphasized in the intervention. Thus, the transition Patterns 112 and 212 (indicating student-level latent Class 1 at the pre-test and student-level latent Class 2 at the post-test) showed an intervention effect. As can be seen in Table 3 (bottom), teacher-level latent classes had different proportions of transition patterns. For teacher-level Class 1, there were more students in student-level Class 1 at both pre- and post-test (i.e., Pattern 111) than in student-level Class 1 in pre-test and student-level Class 2 on post-test (i.e., Pattern 112). Within teacher-level Class 2, however, there were more students who moved from student-level Class 1 on the pre-test to student-level Class 2 on the post-test (i.e., Pattern 212) than student-level Class 1 on both pre- and post-tests (i.e., Pattern 211). Transition Pattern 212 is of interest as it indicates the intervention effect: All teachers in teacher-level latent Class 2 (and their students) were in the EAI intervention group.

Over the course of the host study, trained observers collected a total of 324 whole-class observations, 173 of them in EAI classrooms. A secondary observer was present during 20 % of the observed class periods. Field notes included demographic information (e.g., date, school,

TABLE 5.
Error code classification with proportion by student-level latent classes from a multilevel LTA-MixIRTM.

Items	Group membership	No error	Errors										
			C	AA	SD	AC	EQ	CE	WO	O	L/S	RN	NS
1	1	0.629	0.583	0.115	–	0.031	–	0.010	–	0.260	–	–	–
	2	0.371	0.556	0.111	–	0.000	–	0.111	–	0.222	–	–	–
2	1	0.634	0.577	0.134	0.010	0.072	0.000	0.010	–	0.196	–	–	–
	2	0.366	0.455	0.091	0.091	0.000	0.091	0.000	–	0.273	–	–	–
3	1	0.420	0.534	0.068	0.110	0.037	0.052	0.016	–	0.152	–	–	0.031
	2	**0.580**	0.368	0.026	0.237	0.026	0.079	0.000	–	0.184	–	–	0.079
4	1	0.449	0.521	0.068	0.111	0.026	0.058	0.005	–	0.163	–	–	0.047
	2	**0.551**	0.429	0.024	0.214	0.024	0.048	0.095	–	0.167	–	–	0.000
5	1	0.311	0.485	0.063	0.112	0.034	0.063	0.024	–	0.160	–	–	0.058
	2	**0.689**	0.395	0.023	0.209	0.023	0.093	0.047	–	0.209	–	–	0.000
6	1	0.233	0.469	0.062	0.152	0.038	0.085	0.019	–	0.128	–	–	0.047
	2	**0.767**	0.462	0.026	0.231	0.000	0.103	0.000	–	0.179	–	–	0.000
7	1	0.263	0.481	0.038	0.105	0.029	0.086	0.014	–	0.176	–	–	0.071
	2	**0.737**	0.349	0.023	0.163	0.023	0.140	0.070	–	0.209	–	–	0.023
8	1	0.296	0.478	0.043	0.110	0.033	0.062	0.019	–	0.172	–	–	0.081
	2	**0.704**	0.255	0.021	0.234	0.021	0.149	0.085	–	0.170	–	–	0.064
9	1	0.288	0.452	0.029	0.105	0.033	0.076	0.019	–	0.205	–	–	0.081
	2	**0.712**	0.229	0.042	0.188	0.000	0.188	0.063	–	0.208	–	–	0.083
10	1	0.296	0.474	0.029	0.086	0.043	0.067	0.019	–	0.201	–	–	0.081
	2	**0.704**	0.234	0.043	0.191	0.000	0.085	0.128	–	0.234	–	–	0.085
11	1	0.328	0.498	0.063	0.088	0.015	0.044	0.010	–	0.205	–	–	0.078
	2	**0.672**	0.250	0.045	0.227	0.000	0.182	0.045	–	0.182	–	–	0.068
12	1	0.309	0.481	0.048	0.091	0.014	0.063	0.005	–	0.216	–	–	0.082
	2	**0.691**	0.213	0.043	0.213	0.000	0.170	0.043	–	0.234	–	–	0.085
13	1	0.227	0.409	0.037	0.107	0.014	0.070	0.014	–	0.251	–	–	0.098
	2	**0.773**	0.176	0.039	0.216	0.000	0.098	0.137	–	0.196	–	–	0.137
14	1	0.182	0.433	0.041	0.134	0.018	0.065	0.018	–	0.198	–	–	0.092
	2	**0.818**	0.224	0.041	0.184	0.000	0.061	0.143	–	0.265	–	–	0.082
15	1	0.646	0.464	0.018	0.018	0.018	0.018	0.027	0.045	0.304	–	–	0.089
	2	0.354	0.174	0.043	0.043	0.000	0.000	0.130	0.043	0.261	–	–	0.304
16	1	0.288	0.476	0.005	0.120	–	0.053	0.024	0.014	0.216	–	–	0.091
	2	**0.712**	0.256	0.000	0.279	–	0.000	0.023	0.023	0.209	–	–	0.209
17	1	0.579	0.365	0.006	0.019	–	0.013	0.101	0.050	0.327	–	–	0.119
	2	0.421	0.135	0.027	0.027	–	0.000	0.243	0.054	0.270	–	–	0.243
18	1	0.222	0.018	0.005	0.005	–	0.005	0.005	0.014	0.301	0.502	0.027	0.119
	2	**0.778**	0.016	0.016	0.000	–	0.016	0.016	0.031	0.234	0.500	0.047	0.125
19	1	0.222	0.386	0.009	0.088	0.005	0.074	0.019	0.014	0.298	–	–	0.107
	2	**0.778**	0.160	0.020	0.200	0.000	0.100	0.020	0.060	0.260	–	–	0.180
20	1	0.200	0.324	0.005	0.054	0.005	0.027	0.005	0.014	0.329	0.144	0.000	0.095
	2	**0.800**	0.068	0.014	0.082	0.000	0.055	0.014	0.014	0.219	0.329	0.068	0.137

See Appendix for the description of error types; Irrelevant errors indicated by "–".

condition, unit of instruction), level of treatment fidelity (e.g., surface features, quality of implementation), and descriptions of classroom activities. The content for the observations recording treatment fidelity of EAI was derived from the daily lesson plans teachers were provided.

Overall, teachers followed the lesson plans closely. Observers indicated that EAI teachers taught activities in the warm-up 70 % of the time and the main lesson 95 % of the time. Inter-observer agreement was 71 % for the warm-up and 94 % for main part of the lesson. Observers reported that on several occasions teachers were unable to finish the entire lesson plan in one day for various reasons (e.g., shorter class time, behavioral problems, other school activities) so the lesson was continued on the following day. Discussions with observers at the conclusion of the study indicated that these variations in schedule explained much of the inconsistencies in the way the warm-up activities were implemented (e.g., taught the first day but not the next) and how they were scored.

At the conclusion of the study, observers agreed that all of the EAI teachers displayed good teaching skills over the entire course of the study. During the problem-solving units, observers did notice subtle differences in describing key math concepts and procedures, facilitating student discussions, answering questions, and dealing with the complexity of the applied projects, but there were little or no qualitative differences between teachers during the FAW unit.

The similarity in teaching styles during the FAW unit can be explained by the content of the instructional materials and the way FAW lessons were structured. The multimedia-based lessons contain a teacher voice describing the fractions representations on the screen. The use of manipulatives (e.g., fraction strips) that accompany the lessons is partially scripted for teachers to follow. Therefore, the lessons are explicitly directed, tightly managed, and leave little room for teacher variation. This design element was important in the development of FAW because difficulty in explaining fractions concepts has been an area of concern in teacher education for many years (Ball, 1990). Teaching and student factors that were more subtle than the classroom observers were able to detect likely accounted for the identification of teacher type differences in the multilevel LTA–MixIRTM.

(iii) *How much progress do students make and does the progress depend on the intervention, on the teachers, or on the combination of teacher and intervention*? Unlike the multilevel LTA–LCM (Asparouhov & Muthén, 2008; Yu, 2007), individual differences in ability are allowed in each transition pattern in the multilevel LTA–MixIRTM. Table 6 shows the multivariate ability distribution population parameter estimates for each transition pattern. In the table, the labels are shown as equal as group membership was assumed to be the same at each time point (as illustrated in Figure 1). For example, Patterns 111 and 121 shared the same student-level latent Class 1, in teacher-level Class 1 on the post-test. Thus, these two patterns shared constraints for model identification, $\mu^*_{1_2 1} = 0$ and $\sigma^{2*}_{1_2 1} = 1$.

Larger variances (i.e., larger individual differences in ability) were observed in the larger of the two student-level classes characterizing each of the two teacher-level latent classes. Teacher-level Class 1 consisted of more students clustered into student-level Class 1 (the low-ability group). The remaining students in teacher-level Class 1 were those classified into student-level Class 2 (the high-ability group). The variance of these latter students was small and non-significant, $\widehat{\sigma}^2_{2_2 1} = 0.044$, in both Transition Pattern 112. Similarly, students in student-level latent Class 2 (i.e., the high ability group) comprised the larger of the student-level latent classes in teacher-level Class 2. The students in student-level latent Class 1 (the low-ability group) were in the minority in this teacher-level latent class. Of these latter students, there was only a small and non-significant variance in ability, $\widehat{\sigma}^2_{1_2 2} = 0.012$, for Transition Pattern 211. Apart from the measurement and estimation errors, these results were to be expected given the homogeneity of teacher-level class with respect to student-level latent class.

The pre-test and post-test scores plotted for four representative transition patterns (see Figure 2) clearly indicate that scores of students tended to cluster by transition pattern. Most students' scores in all four patterns were very low on the pre-test, but increased on the post-test.

TABLE 6.
Estimates (standard errors in bracket) for population parameters of multidimensional ability distribution in a multilevel LTA–MixIRTM.

Transition pattern	Time	$\vec{\mu}$	Σ_{g_t}	
			Time 1	Time 2
111	Time 1	$\mu^*_{1_11} = 0$ (1)	$\sigma^{2*}_{1_11} = 1$ (9)	
	Time 2	$\mu^*_{1_21} = 0$ (2)	$\widehat{\sigma}_{111} = 0.856$ [0.037] (14)	$\sigma^{2*}_{1_21} = 1$ (10)
112	Time 1	$\mu^*_{1_11} = 0$ (1)	$\sigma^{2*}_{1_11} = 1$ (9)	
	Time 2	$\widehat{\mu}_{2_21} = 0.351$ [0.154] (4)	$\widehat{\sigma}_{112} = 0.011$ [0.020] (15)	$\widehat{\sigma}^2_{2_21} = 0.044$ [0.090] (12)
211	Time 1	$\widehat{\mu}_{1_12} = -1.167$ [0.153] (5)	$\widehat{\sigma}^2_{1_12} = 0.898$ [0.254] (17)	
	Time 2	$\widehat{\mu}_{1_22} = -0.003$ [0.039] (6)	$\widehat{\sigma}_{211} = 0.004$ [0.009] (22)	$\widehat{\sigma}^2_{1_22} = 0.012$ [0.167] (18)
212	Time 1	$\widehat{\mu}_{1_12} = -1.167$ [0.153] (5)	$\widehat{\sigma}_{1_12} = 0.898$ [0.254] (17)	
	Time 2	$\widehat{\mu}_{2_22} = 0.859$ [0.156] (8)	$\widehat{\sigma}_{212} = 0.212$ [0.132] (23)	$\widehat{\sigma}_{2_22} = 0.833$ [0.293] (20)

$*$: constraints for model identification; $**$: constraints for data structure.
Values in parentheses are labels of constraints presented in Table 1 and Figure 1.
Note. Subscripts for the mean and variance are $g_t h$ where g is a student-level latent class, t is a time point, and h is a teacher-level latent class. The name of "Pattern XYZ" is from the definition that "X" is a teacher-level latent class, "Y" is a student-level latent class at Time 1, and "Z" is a student-level latent class at Time 2.

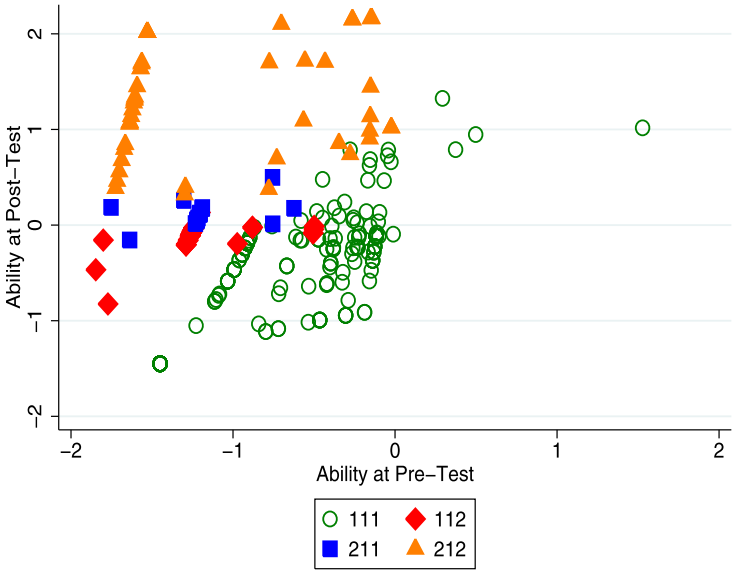

FIGURE 2.
Ability scores at two time points of a multilevel LTA–MixIRTM.

The changes in ability, however, clearly differed by transition patterns. The correlation between pre-test and post-test scores for Pattern 111 was 0.856. This indicates that students in transition Pattern 111, who had high or low scores on the pretest, tended to retain their relative positions within that pattern on the post-test. However, for students in transition Pattern 211, there was a zero correlation ($r = 0.004$) between pre-test and post-test scores. As also shown in Figure 2 (the squares in the figure are for Pattern 211), the reason for this was that there was little variability in performance for these students on post-test (although as noted below, there was a large dif-

ference in average performance between pre-test and post-test). Students who were classified in teacher-level Class 2 (i.e., in Patterns 211 and 212) had higher levels in general on the post-test than those who were classified in teacher-level Class 1. Means and standard deviations of *change* (i.e., post-test score – pre-test score) were 0.138 and 0.365, respectively, for students with Pattern 111, 1.099 and 0.234, respectively, for students with Pattern 112, 1.305 and 0.258, respectively, for students with Pattern 211, and 2.354 and 0.775, respectively, for students with Pattern 212.

5.6. Empirical Study Summary and Discussion

No student-level latent classes were detected in the pre-test based on BIC using the MixIRTM. Most students' scores, in fact, were very low with little variability. The effects of instruction in either the EAI or BAU conditions could be seen in an increase in both the dimensionality and the variability of students' scores in the post-test data. Two student-level latent classes were found in the post-test: student-level Class 1 was lower in ability than student-level Class 2. These student-level latent classes were characterized by different patterns of errors made on the test questions. Examination of these error patterns suggested that members of each student-level latent class differed in the way they solved the problems posed by the test questions.

A significant covariate on transition proportions indicated a significant EAI intervention effect. Following instruction (either EAI or BAU), four transition Patterns (111, 112, 211, and 212) were detected. There were 27.4 % of students with Patterns 112 and 212. Students with these patterns moved from the low ability group to the high ability group. These 27.4 % differed with respect to the type of teachers they had. Students with Pattern 212 were all from the EAI group of teachers who were observed to consistently do a thorough job of implementing the lesson plans. Finally, there were relatively large changes in abilities over time, particularly for students in Patterns 211 and 212, all of whom were from the EAI treatment group.

An interesting artifact in the overall improvement of the EAI students was an increasing error trend by a few students on the post-test who presumably knew they needed to find a new denominator but made mistakes attempting to find it. Students from student-level Class 1 primarily tended to make combining errors, whereas students from student-level Class 2 made less of these but made relatively more errors of the following types instead: selection of one denominator (SD), misrepresentation of equivalent fractions (EQ), and calculation errors (CE) for *like* items. Student-level Class 2 students seemed to realize better that they could not operate independently with numerators and denominators. They also appeared to understand better what fractions were without necessarily seeing all implications, so that they might still make SD and CE errors. We note this as a positive finding because students at least recognized the need for finding common denominators prior to adding or subtracting.

Classroom observations revealed clear differences in the way EAI teachers presented the instructional material and this most likely affected the results. First, observation protocols showed that teachers who taught the lessons with fidelity were more likely to have students making the greater gains. Second, related to our first point, more effective teachers were highly organized and prepared. It was clear they had studied the lesson plans and anticipated ways of answering students' questions. Some teachers had structured their lessons using a schedule like those in the EAI lessons that included a warm-up routine, the presentation of the main lesson, and a summary activity. Finally, roughly one-third of students in the EAI group had a cognitive disability and many of them made remarkable progress in learning complex concepts although not moving up performance levels. Instruction only spanned 18 weeks, which is a relatively brief period of time for low-achieving students to learn how to compute with fractions and improve their problem solving skills.

6. Conclusions and Discussions

LTA–LCM has been suggested as a useful means of investigating stage-sequential change over time and evaluating the effectiveness of intervention, although it has not been used extensively with educational test data. In this study, we used two different LTA models by adding MixIRTMs to the models, an LTA–MixIRTM and a multilevel LTA–MixIRTM. Unlike multilevel LTA–LCM, the individual differences in abilities are modeled within a transition pattern in both the LTA–MixIRTM and the multilevel LTA–MixIRTM. The two models were used to analyze data from a curriculum-based test of mathematics achievement in order to determine the effects of an instructional intervention. The addition of the multilevel data to the LTA–MixIRTM enables researchers to consider both the potential heterogeneity in student response to instructional intervention as well as a methodology for assessing the effects of the intervention over time at both student and teacher levels.

Change is often measured based on observed total scores or (arcsine–square-root transformed) proportions correct for binary responses and mean changes over time using repeated measures analysis of variance (ANOVA) or multivariate analysis of variance (MANOVA) in instructional intervention studies (e.g., Bottge et al., 2007). Maxwell and Tiberio (2007) discuss limitations of using these approaches to measure change. First, the repeated measures ANOVA and MANOVA comparison structures are based on mean change of a total score as a manifest variable across time points and individual differences in change are modeled with an error term. Second, repeated measures ANOVA is based on the strong assumption of sphericity, which implies that every measure must have the same variance and all correlations between any pair of measures must be equal, when there are three or more levels (i.e., time points) of a repeated measure. This assumption of sphericity is likely to be violated in instructional intervention studies due to floor effects before intervention. However, the application of the multilevel LTA–MixIRTM showed that there were individual differences in student growth as a result of the EAI intervention which were not evident using repeated measures ANOVA or MANOVA.

A dimensionality analysis using a MixIRTM indicated that the pre-test data were unidimensional (i.e., no student-level latent classes) and the post-test data were multidimensional (i.e., 2 student-level latent classes having different item parameter estimates). This is likely a common phenomenon in an instructional intervention study when participants have yet to be introduced to the material. The floor effect and its consequence of unidimensionality in the pre-test data lead to a unique transition modeling as all possible transition patterns did not occur in the data. As a result, some possible transition patterns could not be modeled. In the application in this study, only four of the eight possible transition patterns were observed, as no students were detected in student-level Class 2 (labeled the high ability group) on the pre-test. The application is still an illustration of the LTA because transition probabilities can be differentiated from the student-level Class 2 proportions. Further, the illustration showed how the multilevel LTA–MixIRTM could be used for this type of data structure, a structure that is likely common in intervention studies. The model described in the paper also can be used in other applications having full transition patterns.

As noted earlier, multidimensional IRT model and MixIRTM are two conceptually different approaches in IRT for modeling of multidimensionality in test data (Rijmen & De Boeck, 2005). Multidimensional IRT models can be useful, such as when IRT scaled scores for sub-domains are needed for detecting intervention effects (see, e.g., Cho, Athay, & Preacher, 2012a). In the application in this study, the mixture portion of the model was shown to be appropriate for the group-comparison design. Using the post-test results, the fit of the multilevel 2-dimensional 2PL IRT model (with two student-level dimensions and two teacher-level dimensions) was as good as that of the multilevel MixIRTM based (BIC = 3350.248 for the multilevel 2-dimensional 2PL IRT model; BIC = 3353.560 for the multilevel MixIRTM). The multilevel MixIRTM was

chosen as the measurement model for this application for two main reasons. First, in this study research Questions 1 and 2 were more appropriately addressed with a model that enabled a group-comparison design. With respect to Question 1, determining if students engaged differently in problem solving was addressed in this study by analyzing the data to determine whether there were latent subgroups of students who were homogeneous in their responses to items. Similarly, for Question 2, the group-comparison design permitted determining whether students differed with respect to the different latent subgroups of teachers. Second, there was multimodality in the ability score distributions at both the student and teacher levels, respectively, based on the multilevel 2-dimensional 2PL IRT model. The modality does not stem from the difference between BAU and EAI groups because there is also multimodality within each group. This can be taken as evidence supporting the use of mixture modeling.

Measurement invariance over time points is necessary to be able to measure change. The traditional method for checking measurement invariance under an IRT framework is to obtain item parameter estimates using either separate or concurrent estimation. Then a test is conducted to determine whether or not an item functions differently across time points. In the present study, there was a floor effect that resulted in most students having a minimum score in the pre-test: (51 % of students had scores of 0 in the BAU group, and 60 % of students has scores of 0 in the EAI group). In this study, item parameter estimates were obtained using a MixIRTM on the post-test data. These parameter estimates were then used for modeling the pre-test data. The rationale for this was that there were two dimensions over time realized as two latent classes and no individual differences in one of the two dimensions in the pre-test data. This application can be considered as a longitudinal extension of one- and two-dimensional MixIRTMs (De Boeck, Cho, & Wilson, 2011).

The number of student-level and teacher-level latent classes were selected based on model-fit using BIC. Two student-level latent classes were detected by the MixIRTM. These were characterized by different item profiles. Error analysis showed that each student-level latent class had different patterns of errors, indicating different reasons why students answered incorrectly. One item was found to be class invariant and was used as an anchor between the two student-level latent classes. This enabled at least minimal scale comparability across latent classes (Paek, Cho, & Cohen, 2010). Two teacher-level latent classes were mainly characterized by the different proportions of student-level latent classes. Even though the number of latent classes we found were well-explained by the empirical evidence, it was still not clear whether within-class abilities were appropriately compared between classes or whether it is necessary to adjust scores for deficiencies in skill states (Embretson & Reise, 2000).

The same items were used for both the pre- and post-tests. In such a case, it is possible that some memory effects may have played a part in responses for some students. Memory effects, response consistency effects, and practice effects are all problems that can potentially exist when dealing with repeated measures, possibly resulting in violation of local independence within each pattern. A mitigating factor is that in this study 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 pre-tests or post-tests, 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.

Maximum likelihood estimation of model parameters in each latent class was difficult given the small sample sizes. The multilevel LTA–MixIRTM did moderate the data sparseness better, however, than the multilevel LTA–LCM. Estimation of the multilevel LTA–MixIRTM does present a computational problem, as high dimensional integration is required as the number of time points increases. Estimation for our empirical study required 14 hours to complete on a computer equipped with a 3.19-GHz processor with 3.00 GB of RAM with multiple initial values to monitor local maxima.

Acknowledgements

The authors wish to thank Dr. Paul De Boeck (Ohio State University) for his valuable comments on previous drafts of this paper. We also would like to thank three anonymous reviewers and the editor, Dr. Maydeu-Olivares, for their helpful comments and suggestions on previous versions of this paper.

The research reported in this paper was supported by a grant from the U.S. Department of Education, Institute of Education Sciences, PR Number H324A090179. Any opinions, findings, or conclusions are those of the authors and do not necessarily reflect the views of the supporting agency.

Appendix: Error Codes and Examples from a Fractions Computation Test

• *Combining* (*C*): Student combines numerators and combines denominators, consistently applying the same operation to numerator and denominator.

Examples: $\frac{1}{3}+\frac{1}{3}=\frac{2}{6}$, $\frac{7}{8}-\frac{1}{4}=\frac{6}{4}$

• *Add All* (*AA*): Student separately adds together all the components of the fractions.

Example: $\frac{3}{4}+\frac{2}{5}=14$

• *Select Denominator* (*SD*): Student selects one of the denominators listed in the problem and makes no attempt to make equivalent fraction. Denominator given in the answer must be present in the problem.

Examples: $\frac{1}{2}+\frac{3}{16}=\frac{4}{16}$

• *Adding Components* (*AC*): Students adds the numerator and denominator of each individual fraction together and those two sums are represented in the answer.

Example: $\frac{1}{2}+\frac{3}{16}=\frac{3}{19}$

• *Equivalent Fraction Error* (*EQ*): Student makes an error when attempting to represent an equivalent fraction.

• *Computation Error* (*CE*): Student makes an arithmetic error.

• *Wrong Operation* (*WO*): Student adds given a subtraction problem or subtracts given an addition problem.

Example: $3\frac{5}{8}+2\frac{1}{2}=6\frac{1}{8}$

• *Other* (*O*): Student makes error other than those listed above.

• *Large–small* (*L/S*): Student subtracts smaller from larger fraction out of order. Applies only to Fraction Computation Test items 18 and 20. Or student subtracts smaller part of fraction from larger part of fraction out of order when combined with (C) error.

Examples: $7\frac{1}{4}-3\frac{1}{2}=4\frac{1}{4}$, $4-2\frac{3}{4}=2\frac{3}{4}$

• *Renaming* (*RN*): Student makes a mistake when renaming a whole number as a mixed number; the student fails to borrow correctly from a whole number.

• *No response* (*NR*): Student leaves problem blank.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Caski (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Academiai Kiado.

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

Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G.R. Hancock & K.M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27–51). Charlotte: Information Age Publishing

Ball, T.L. (1990). Prospective elementary and secondary teachers understanding of division. *Journal for Research in Mathematics Education*, *21*, 132–144.

Bottge, B.A. (1999). Effects of contextualized math instruction on problem solving of average and below-average achieving students. *Journal of Special Education*, *33*, 81–92.

Bottge, B.A., Heinrichs, M., Mehta, Z., & Hung, Y. (2002). Weighing the benefits of anchored math instruction for students with disabilities in general education classes. *Journal of Special Education*, *35*, 186–200.
Bottge, B.A., Heinrichs, M., Mehta, Z.D., Rueda, E., Hung, Y., & Danneker, J. (2004). Teaching mathematical problem solving to middle school students in math, technology education, and special education classrooms. *RMLE Online*, *27*. http://www.nmsa.org/portals/0/pdf/publications/RMLE/rmle_vol27_no1_article1.pdf. Retrieved July 20, 2006.
Bottge, B.A., Rueda, E., Serlin, R.C., Hung, Y., & Kwon, J. (2007). Shrinking achievement differences with anchored math problems: challenges and possibilities. *Journal of Special Education*, *41*, 31–49.
Bottge, B., Ma, X., Toland, M., Gassaway, L., & Butler, M. (2012). *Effects of enhanced anchored instruction on middle school students with disabilities in math*. Department of Special Education & Rehabilitation Counseling, University of Kentucky, Lexington, KY.
Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.
Cho, S.-J., & Cohen, A.S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*, 336–370.
Cho, S.-J., Cohen, A.S., Kim, S.-H., & Bottge, B. (2010). Latent transition analysis with a mixture IRT measurement model. *Applied Psychological Measurement*, *34*, 583–604.
Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis*, *55*, 12–25.
Cho, S.-J., Athay, M., & Preacher, K. J. (2012a). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical and Statistical Psychology*. doi:10.1111/j.2044-8317.2012.02058.x.
Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2012b). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*. doi:10.1080/00949655.2011.603090.
Cognition Technology Group at Vanderbilt (1990). Anchored instruction and its relationship to situated cognition. *Educational Researcher*, *19*, 2–10.
Cognition Technology Group at Vanderbilt (1997). *The Jasper project: lessons in curriculum, instruction, assessment, and professional development*. Mahwah: Erlbaum.
Collins, L.M., & Wugalter, S.E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*, 131–157.
CCSSI-M (Common Core State Standards Initiative—Mathematics) (2011). http://www.corestandards.org. Retrieved December 5, 2011.
Congdon, P. (2003). *Applied Bayesian modeling*. New York: Wiley.
De Boeck, P., Wilson, M., & Acton, G.S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review*, *112*, 129–158.
De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent DIF. *Applied Psychological Measurement*, *35*, 583–603.
Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence-Erlbaum.
Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.
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: Sage Publications.
Henry, K.L., & Muthén, B. (2010). Multilevel latent class analysis: an application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling. A Multidisciplinary Journal*, *17*, 193–215.
Hickey, D.T., Moore, A.L., & Pellegrino, J.W. (2001). The motivational and academic consequences of elementary mathematics environments: do constructivist innovations and reforms make a difference? *American Educational Research Journal*, *38*, 611–652.
Lee van Horn, M., Fagan, A.A., Jaki, T., Brown, E.C., Hawkins, D., Arthur, M.W., Abbott, R.D., & Catalano, R.F. (2008). Using multilevel mixtures to evaluate intervention effects in group randomized trials. *Multivariate Behavioral Research*, *43*, 289–326.
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.
Maxwell, S.E., & Tiberio, S. (2007). Multilevel models of change: fundamental concepts and relationships to mixed models and latent growth-curve models. In A. Ong & M.H.M. van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (pp. 439–452). New York: Oxford University Press.
McDonald, R.P. (1967). *Psychometric monographs: Vol. 15. Nonlinear factor analysis*. Richmond: Psychometric Corporation. Retrieved from http://www.psychometrika.org/journal/online/MN15.pdf.
McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
Muthén, B., Brown, C.H., Jo, B., Khoo, S.-T., Yang, C.C., Wang, C.-P., Kellam, S.G., Carlin, J.B., & Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics*, *3*, 459–475.
Muthén, B., & Asparouhov, T. (2006). Item response mixture modeling: application to tobacco dependence criteria. *Addictive Behaviors*, *31*, 1050–1066.
Muthén, L.K., & Muthén, B.O. (2006–2010). *Mplus [computer program]*. Los Angeles: Muthén & Muthén.
Nylund, K.L., Asparouhov, T., & Muthén, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535–569.
Paek, I., Cho, S.-J., & Cohen, A.S. (2010). *A note on scaling linking in mixture IRT model applications*. Paper presented at the National Council on Measurement in Education. Denver.

Raudenbush, S.W., & Bryk, A.G. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.
Rijmen, F., & De Boeck, P. (2005). A relation between a between-item multidimensional IRT model and the mixture-Rasch model. *Psychometrika*, *70*, 481–496.
Rijmen, F., De Boeck, P., & van der Maas, H.L.J. (2005). An IRT model with a parameter-driven process for change. *Psychometrika*, *70*, 651–669.
Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.
Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal and structural equation models*. Boca Raton: Chapman & Hall.
Steinley, D., & McDonald, R.P. (2007). Examining factor score distributions to determine the nature of latent spaces. *Multivariate Behavioral Research*, *42*, 133–156.
Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion of model fitting. *Suri-Kagaku (Mathematic Science)*, *153*, 12–18 (in Japanese).
Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*, 213–239.
Vermunt, J.K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*, 33–51.
Vermunt, J.K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: basic and advanced*. Belmont: Statistical Innovations.
von Davier, M., Xu, X., & Carstensen, C.H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318–336.
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.
Yang, C. (2006). Evaluating latent class analyses in qualitative phenotype identification. *Computational Statistics & Data Analysis*, *50*, 1090–1104.
Yu, H.-T. (2007). *Multilevel latent Markov models for nested longitudinal discrete data*. Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.

Manuscript Received: 12 FEB 2012
Final Version Received: 22 JUN 2012
Published Online Date: 5 JAN 2013