Journal of Educational and Behavioral Statistics
2020, Vol. 45, No. 3, pp. 339–368
DOI: 10.3102/1076998619882026
Article reuse guidelines: sagepub.com/journals-permissions

On Longitudinal Item Response Theory Models: A Didactic

Chun Wang
University of Washington

Steven W. Nydick
Korn Ferry

Recent work on measuring growth with categorical outcome variables has combined the item response theory (IRT) measurement model with the latent growth curve model and extended the assessment of growth to multidimensional IRT models and higher order IRT models. However, there is a lack of synthetic studies that clearly evaluate the strength and limitations of different multilevel IRT models for measuring growth. This study aims to introduce the various longitudinal IRT models, including the longitudinal unidimensional IRT model, longitudinal multidimensional IRT model, and longitudinal higher order IRT model, which cover a broad range of applications in education and social science. Following a comparison of the parameterizations, identification constraints, strengths, and weaknesses of the different models, a real data example is provided to illustrate the application of different longitudinal IRT models to model students' growth trajectories on multiple latent abilities.

Keywords: *item response theory; latent growth curve model; overall ability; domain ability*

1. Introduction

In education, one is often interested in determining student growth. These changes can sometimes be captured by latent variable models. The latent variables, such as students' abilities, are typically measured by binary (or polytomous) responses to items. Item response theory (IRT) models are useful tools to model the relationship between the categorical outcome variables and the latent continuous traits. Recent work has extended IRT models to model changes in latent traits, leading to the family of longitudinal IRT (L-IRT) models (e.g., Andersen, 1985; Cai, 2010; Hsieh, von Eye, & Maier, 2010; Huang, 2013; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009; Paek, Li, & Park, 2016; von Davier, Xu, & Carstensen, 2011; Wang, Kohli, & Henn, 2016; Wilson,

Zheng, & McGuire, 2012). Within this family, models differ mainly in the following aspects: (1) the measurement model that implies the factor structure of the primary latent traits measured repeatedly, which could either be unidimensional, multidimensional (Hsieh et al., 2010), or hierarchical (Huang, 2013); (2) the relationship of the latent traits over time, which could either be captured by a completely unstructured covariance matrix (Andrade & Tavares, 2005; Cai, 2010; Paek et al., 2016) or by linear/nonlinear change patterns via the latent growth curve (LGC) models (Bollen & Curran, 2006; Duncan, Duncan, & Strycker, 2006); and (3) whether nuisance factors are in place to account for the dependency of the same items administered over time (e.g., two-tier model; Cai, 2010; Paek et al., 2016; Wang et al., 2016).

Due to the well-known connection between IRT and categorical factor analysis (e.g., Takane & de Leeuw, 1987), L-IRT models can also be discussed in structural equation modeling (SEM) terms. However, IRT offers two conceptual advantages: (1) assuming item (or anchor item) parameters are the same over time to ensure longitudinal invariance of the lowest order traits and (2) incorporating guessing parameters into the functional form of the model.

Different forms of L-IRT models were proposed by different groups of researchers, and they have all been individually demonstrated to work well; however, few studies have explored the connections among the models or the strengths and limitations of each of them. Our goal here is to capitalize on the shared features and distinctions among various L-IRT models to provide practitioners with coherent guidelines about the conditions under which each model could be applied and/or should be preferred.

Three specific types of models will be the focus of discussion. In order of complexity, these models include the longitudinal unidimensional IRT (L-UIRT) model (Wang et al., 2016; Wilson et al., 2012), longitudinal multidimensional IRT (L-MIRT) model (Hsieh et al., 2010), and longitudinal higher-order IRT (L-HO-IRT) model (Huang, 2013). All of these models are variations of the general LGC model and the respective measurement model: The UIRT model assumes that a single latent trait is measured by all the items; MIRT models posit that item responses are probabilistically determined by multiple, usually correlated, latent traits; the HO-IRT models (de la Torre & Song, 2009; Sheng & Wikle, 2008) capture the hierarchical nature of factor structure (e.g., Huang & Wang, 2014; Sawaki, Stricker, & Oranje, 2009), whereby a general factor (such as math aptitude) informs domain-specific factors (such as algebra, geometry, calculus, or subsets thereof). These three models were selected to cover a majority of practical applications. Moreover, LGC models were chosen over an unstructured covariance matrix because LGC results in both group-level and individual-level growth trajectories, which are often useful for interpreting data patterns. On the other hand, LGC introduces additional latent variables (i.e., individual intercepts and slopes) that complicate model identification constraints and requires additional guidelines for model estimation. Note that the L-MIRT model with

unstructured covaraince matrix of θ over time is discussed in detail in Paek, Li, and Park (2016).

In the remaining sections, we introduce the three models and explain when each model could be applied. For each model, we describe identification constraints, which can be different depending on whether some items have pre-calibrated parameters. After determining the identification requirements, we are then ready to estimate the models. Estimation presents various challenges, and we describe the available estimation methods, complications due to high dimensionality, and possible solutions. We finally illustrate the models with a real data example.

2. L-IRT Models

2.1. L-UIRT Model

If only one primary latent trait is measured over time, then the simplest model, the L-UIRT model, can be applied. Let $\boldsymbol{\theta}_i$ denote the T-by-1 vector of the unidimensional trait for person i across T time points. Assume there are p fixed (denoted as $\boldsymbol{\beta}$) and q random (denoted as $\boldsymbol{\nu}_i$) effects explaining the growth pattern of θ. Then, the LGC model on $\boldsymbol{\theta}_i$ can be written in a general form as follows:

$$\boldsymbol{\theta}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu}_i + \boldsymbol{\delta}_i. \tag{1}$$

In Equation 1, $\mathbf{X}$ and $\mathbf{Z}$ are the T-by-p and T-by-q design matrices for the fixed effects and random effects, respectively. In a simple LGC model with only random intercepts and random slopes, $p = q = 2$ and $\mathbf{X} = \mathbf{Z} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & T-1 \end{pmatrix}$. $\boldsymbol{\delta}_i$ is a T-by-1 vector of residuals. The random effects are often assumed to follow a multivariate normal distribution with a mean of 0s and a covariance matrix of Σ_ν. Note the number of measurement occasions, T, can be different for each person in the LGC model, allowing for missing data by design. For simplicity, we keep T the same across persons in this article.

For a simple linear growth model with a single person-specific intercept and slope, we can rewrite Equation 1 as

$$\theta_i^t = \pi_{0i} + \pi_{1i} \times (t-1) + \delta_i^t, \tag{2}$$

where π_{0i} and π_{1i} are the individual intercept and slope parameters. The individual intercepts/slopes can be further written as deviations from an overall intercept (β_0) and slope (β_1) as $\pi_{0i} = \beta_0 + \nu_{0i}$ and $\pi_{1i} = \beta_1 + \nu_{1i}$.

The latent variable described by Equation 2, θ_i^t, can be measured by responses to assessment items. Assuming that responses are binary, one can model the

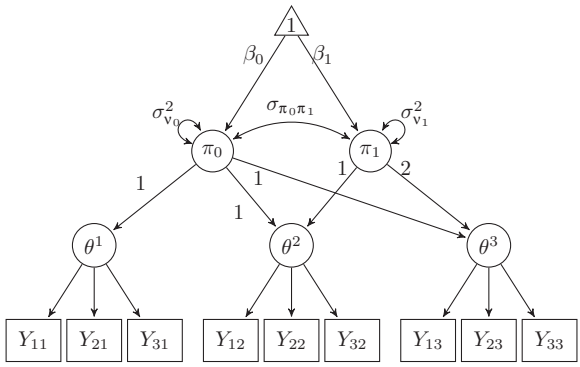

FIGURE 1. *A path diagram for the longitudinal unidimensional item response theory model with three items per time point and three time points. π_0 represents the random intercept parameter per person, whereas π_1 represents the random linear slope parameter per person. β_0 and β_1 are the population means of π_0 and π_1, respectively.*

probability of correctly responding to any item given a particular value of the latent variable with the two-parameter logistic (2PL) model. The 2PL defines the probability of examinee i correctly responding to item j by the following item response function (IRF):

$$P_j(\theta_i^t) = \Pr\left(Y_{ij}^t = 1|\theta_i^t, a_j^t, b_j^t\right) = \frac{1}{1 + \exp\left[-a_j^t\left(\theta_i^t - b_j^t\right)\right]}, \tag{3}$$

where a_j^t and b_j^t refer to discrimination and difficulty parameters for item j administered at time t. This notation is flexible enough to accommodate item sets varying across time. Figure 1 shows an illustrative path diagram of the L-UIRT model with three hypothetical time points and three items per time point.

Many large-scale educational surveys have primary measurements that differ from one occasion to another (Edwards & Wirth, 2009; McArdle et al., 2009). Yet, to establish a common scale, one must either have a common set of anchor items that is shared across time or sets of anchor items that already have parameters precalibrated and put on a common scale (e.g., Wang et al., 2016). Kolen and Brennan (2004) recommended that assessments should have at least 20% of items to anchor the parameters to the common scale. If enough items are linked across time, and assuming no item parameter drift, then assessments with unknown item parameters require some model identifiability constraints to be imposed. Constraints are required to fix the mean and variance of the latent variable (ξ) at one time point (commonly $t = 1$). Given this constraint, the scale of ξ at the remaining time points will then be determined through the linking items. These constraints include:

1. All of the residuals having mean 0 (i.e., $E(\delta_i^t) = 0$ for all $t = 1, \ldots, T$). This is a typical assumption in parametric regression analysis.
2. The mean of the person-specific intercept parameter being set to 0 (i.e., $\mu_{\pi_{0i}} = \beta_0 = 0$). The purpose of this assumption is to fix the mean of θ at $t = 1$ to 0.
3. The residual variance at the first time point being fixed to be a constant (i.e., $\sigma^2_{\delta_i^{(1)}} = c_1$, where c_1 is some specified constraint). This constraint indirectly fixes the variance of θ at $t = 1$.

Note that after imposing a growth curve structure on θ, θ becomes an endogenous variable in Equations 1 and 2. Hence, instead of directly fixing the mean and variance of θ (as is often desired), most SEM software packages (such as M*plus*) only allow fixing its intercept and the residual variance. The value of c_1 is arbitrary and results in the variance of θ at $t = 1$ becoming the sum of the intercept variance $\left(\text{i.e.,}[\Sigma_u]_{(1,1)}\right)$ and c_1. When anchor items are precalibrated with known parameters, then only the first constraint is necessary to identify the model.

2.2. L-MIRT Model

As a multivariate extension of the L-UIRT model, the L-MIRT model combines the MIRT model with the associative LGC model. The earliest version of the L-MIRT model was proposed by McArdle (1988) and called the "curve of factors" (CUFFS) model. The CUFFS model was developed for multiple, correlated latent traits being tracked over time. For instance, the National Educational Longitudinal Study (NELS: 88) tracked students' academic performance across three measurement occasions on four correlated cognitive scales: mathematics, reading, science, and social studies. In this case, the L-MIRT instead of L-UIRT can better recover the group-level and individual-level growth trajectories by considering all related information. Please note that name "L-MIRT" instead of "CUFFS" is used throughout the didactic for consistency with the other models' names.

Let $\boldsymbol{\theta}_i = (\theta_{i1}^1, \ldots, \theta_{iK}^1, \ldots, \theta_{i1}^T, \ldots, \theta_{iK}^T)'$ be a $KT \times 1$ vector, where T denotes the number of time points and K denotes the number of correlated latent traits (i.e., dimensions) measured at each time point. Assume again that there are p fixed and q random effects per dimension. Then, the general multivariate LGC model can be written as

$$\boldsymbol{\theta}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu}_i + \boldsymbol{\delta}_i. \tag{4}$$

Similar to the notations in Equation 1, $\mathbf{X}$ and $\mathbf{Z}$ are the $KT \times Kp$ and $KT \times Kq$ design matrices. The fixed effect, $\boldsymbol{\beta}$, is a $Kp \times 1$ vector, which is arranged in the following order: (1) the K intercepts, (2) the K slopes for the first fixed covariate, (3) the K slopes for the second fixed covariate, and so on, until (p) the K slopes for the $(p-1)$ th fixed covariate. This can be written in an equation as $\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, \ldots, \beta_{0K}, \beta_{11}, \beta_{12}, \ldots, \beta_{1k}, \ldots, \beta_{(p-1)1}, \ldots, \beta_{(p-1)K})'$.

Similarly, $\boldsymbol{\nu}_i$ is a $Kp \times 1$ vector of random effects with a covariance matrix represented by Σ_ν. Often, Σ_ν is assumed to be a full matrix, which allows random intercepts and slopes to be correlated within and across all domains. Finally, the residuals of $\boldsymbol{\theta}_i$ are represented by $\boldsymbol{\delta}_i = (\delta_{i1}^1, \ldots, \delta_{iK}^1, \ldots, \delta_{i1}^T, \ldots, \delta_{iK}^T)'$, a $KT \times 1$ random vector. The covariance matrix of $\boldsymbol{\delta}_i$, Σ_δ, is often assumed to be diagonal and have the following structure:

$$\begin{pmatrix} \Sigma_1 & \cdots & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \Sigma_T \end{pmatrix}_{KT \times KT},$$

where $\Sigma_t = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2)$ and Σ_δ has T such diagonal blocks.

To be consistent with the description of the L-UIRT model, assume that each domain-level latent trait follows a simple linear trajectory without any additional covariates, which is analogous to the assumption made in the preceeding section. Then $p = q = 2$. If $T = 4$, then both $\mathbf{X}$ and $\mathbf{Z}$ take the form of $\begin{pmatrix} I_K & 0I_K \\ I_K & 1I_K \\ I_K & 2I_K \\ I_K & 3I_K \end{pmatrix}$, where I_K is the $K \times K$ identity matrix. If nonlinear growth trajectories are considered, such as a quadratic effect of time, then $\mathbf{X}$ and $\mathbf{Z}$ would need to be updated with additional columns to account for these effects.

We can also rewrite the model by expanding Equation 4 as follows:

$$\theta_{ik}^t = \pi_{i0k} + \pi_{i1k} \times (t-1) + \delta_{ik}^t, \tag{5}$$

where π_{i0k} and π_{i1k} denote the individual intercept and slope parameters for person i on domain k. As before, the individual intercepts/slopes can be further written as deviations from an overall intercept on domain k (β_{0k}) and slope on domain k (β_{1k}), or

$$\pi_{i0k} = \beta_{0k} + \nu_{i0k}, \tag{6}$$

$$\pi_{i1k} = \beta_{1k} + \nu_{i1k}. \tag{7}$$

The L-MIRT IRF takes the form of

$$P_j(\boldsymbol{\theta}_i^t) = \Pr\left(Y_{ij}^t = 1 | \boldsymbol{\theta}_i^t, a_j^t, b_j^t\right) = \frac{1}{1 + \exp\left[-(\mathbf{a}_j^t)^{\mathrm{T}} \boldsymbol{\theta}_i^t + b_j^t\right]}, \tag{8}$$

where $\mathbf{a}_j^t$ is a vector of discrimination parameters for item j at time t, and "T" denotes transpose. This equation is general enough to include both within-item and between-item multidimensionality structures (Recakase, 2009). Figure 2 provides an illustrative path diagram for a L-MIRT model with

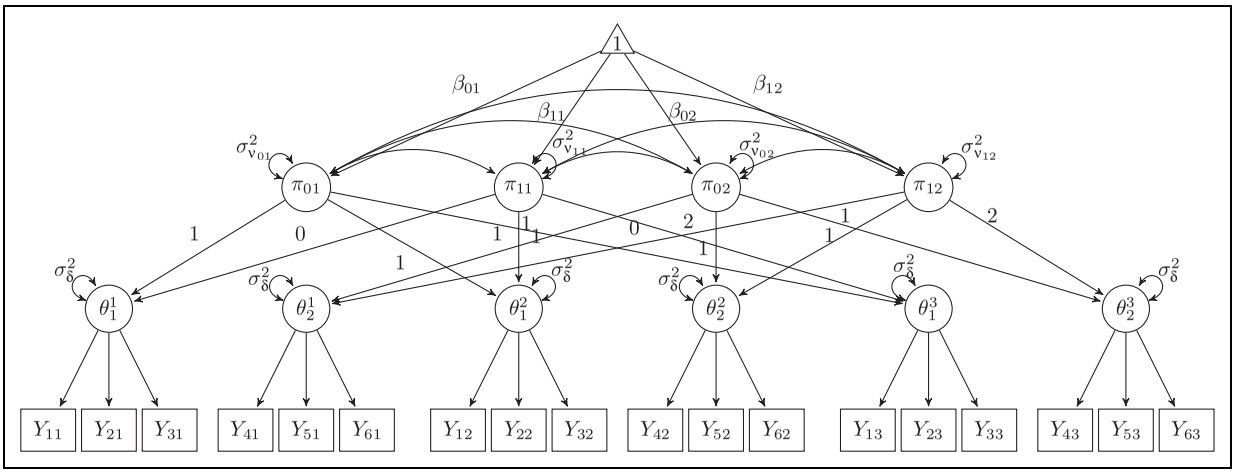

FIGURE 2. *A path diagram for the longitudinal multidimensional item response theory model with three items per domain-level trait, two domain-level traits per time point, and three time points.* π_{01} *and* π_{02} *represent the random intercept parameters per person for both domains, whereas* π_{11} *and* π_{12} *represent the random linear slope parameters per person.* β_{01}, β_{12}, β_{11}, *and* β_{12} *are the population means of* π_{01}, π_{02}, π_{11}, *and* π_{12}, *respectively.*

three measurement occasions, two domains per measurement occasions, and three items per domain. This path diagram only illustrates between-item multidimensionality.

As in the L-UIRT model, items can differ across time, as reflected by the superscript t on item parameters in Equation 8, but anchor items must still be embedded in the item parameter sets to link the scale. Because each domain has a potentially unique scale, anchor items must load on every domain, so that the scale of θ_{ik} is linked across time for all $k = 1, \ldots, K$. As in the unidimensional case, if enough items are linked across time but all item parameters are unknown, then constraints are required to determine the scale of θ_{ik} for $k = 1, \ldots, K$. These constraints are similar to those for the L-UIRT model and include

1. All of the residuals having mean 0 (i.e., $E(\delta_{ik}^{t}) = 0$ for all $t = 1, \ldots, T$ and $k = 1, \ldots, K$).
2. The mean of the person-specific intercept parameters being set to 0 (i.e., $\mu_{\pi_{0ik}} = \beta_{0k} = 0$ for all $k = 1, \ldots, K$). The purpose of this assumption is to fix the mean of θ_{ik}^{t} at $t = 1$ to 0 for all $k = 1, \ldots, K$.
3. The residual variances at the first time point being set to a constant (i.e., $\sigma^2_{\delta_{ik}^{1}} = c_{1k}$, $k = 1, \ldots, K$). As in the unidimensional case, fixing the variance of θ_{ik}^{t} at $t = 1$ (for all k) fixes the variances of θ_{ik}^{t} for the remaining time via the linking items. Moreover, θ_{ik}^{t} is endogenous to the model, so that the variance of θ_{ik}^{t} can only be constrained via its residual variance after partialing out the exogenous fixed and random effects.

As before, when anchor items are precalibrated with known parameters, only the first constraint must be specified to identify the model.

2.3. L-HO-IRT Model

Hierarchical factor structures often emerge in the social sciences to represent a latent construct of interest such as intelligence (Golay & Lecerf, 2011), cognitive ability (Murray & Johnson, 2013), or personality (DeYoung, 2006). General factors are often comprised of several highly related specific factors (a.k.a. first-order factors), each of which is measured by multiple indicators (usually referred to as items). For example, in many educational assessments, one is often required to report both overall proficiency for accountability purposes as well as domain-specific proficiency for diagnostic purposes. To this end, the HO-IRT model was developed by introducing a higher order ability (de la Torre & Hong, 2010; de la Torre & Song, 2009) that relates to each of the first-order abilities. The HO-IRT model contains two levels: (1) a link between a single overall latent trait and one of several domain latent traits and (2) a probabilistic relationship between each domain latent trait and items designed to measure that domain. Specifically, let θ represent the domain latent trait underlying responses to test items and denote ξ as the higher order trait. Then, one can hypothesize that

$$\theta_{ik} = \lambda_k \xi_i + \epsilon_{ik}, \tag{9}$$

where ξ_i is the overall ability of examinee i, θ_{ik} represents domain-specific ability $k \in \{1, \ldots, K\}$ for examinee i, λ_k indicates the relationship between domain-specific ability k and overall ability, and ϵ_{ik} is a disturbance term that can be interpreted as the domain-specific component of the ability not explainable by ξ_i. According to de la Torre and Song (2009), the residuals in Equation 9 are usually assumed uncorrelated across domains, which results in $\boldsymbol{\epsilon}_i$ (containing all of the ϵ_{ik}s) having a diagonal covariance matrix. Note that the variance of ϵ_{ik} is the unique variance of the first-order factor that is not shared by the common second-order factor. At a lower level, the probability of examinee i correctly responding to item j on domain k is defined by the same IRF in Equation 3 except replacing ξ_i with θ_{ik}. As a result, the IRF in Equation 3 implies between-item multidimensionality that is often assumed in the HO-IRT models (e.g., de la Torre & Song, 2009; Wang, 2014). Other measurement models could also be considered based on the properties of the test.

To extend the HO-IRT model across T time points, assume the second-order factor (i.e., overall ability) follows the LGC model, as in Equation 1. Then, the domain-specific ability for person i at time t would also be predicted to systematically change over time (Huang, 2013, 2015) as follows:

$$\boldsymbol{\theta}_i = \boldsymbol{\lambda}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu}_i + \boldsymbol{\delta}_i) + \boldsymbol{\epsilon}_i. \tag{10}$$

Equation 10 can be further understood by expanding it using a scalar equation. That is, given Equations 6 and 7, a domain-specific ability for person i at time point t, θ_{ik}^t, would also follow a linear change over time,

$$\begin{aligned}
\theta_{ik}^t &= \lambda_k \xi_i^t + \epsilon_{ik}^t = \lambda_k(\pi_{0i} + \pi_{1i} \times (t-1) + \delta_i^t) + \epsilon_{ik}^t, \\
&= \lambda_k \pi_{0i} + \lambda_k \pi_{1i} \times (t-1) + (\lambda_k \delta_i^t + \epsilon_{ik}^t), \\
&= \zeta_{0ki} + \zeta_{1ki} \times (t-1) + \upsilon_{ik}^t.
\end{aligned} \tag{11}$$

Notably, Equation 11 implies that the loading of the domain-specific factors on the overall factor remains the same over time, as indicated by the lack of a superscript t on λ_k. By assuming invariance of the factor structure, Equation 11 ensures that the lower order factors carry the same meaning over time, which fulfills the "longitudinal measurement invariance" property (Chen, Sousa, & West, 2006; Liu et al., 2017). Figure 3 provides an illustrative path diagram of the L-HO-IRT model, assuming three time points, two domain-specific abilities per time point, and three items measuring each domain-specific ability.

As shown in Equations 5 and 11, the L-HO-IRT model is nested within the L-MIRT model. This is because the L-MIRT model allows for separate, potentially unrelated, individual intercept and slope parameters across each dimension (i.e., π_{i0k} and π_{i1k}). Conversely, the L-HO-IRT model restricts the domain-level

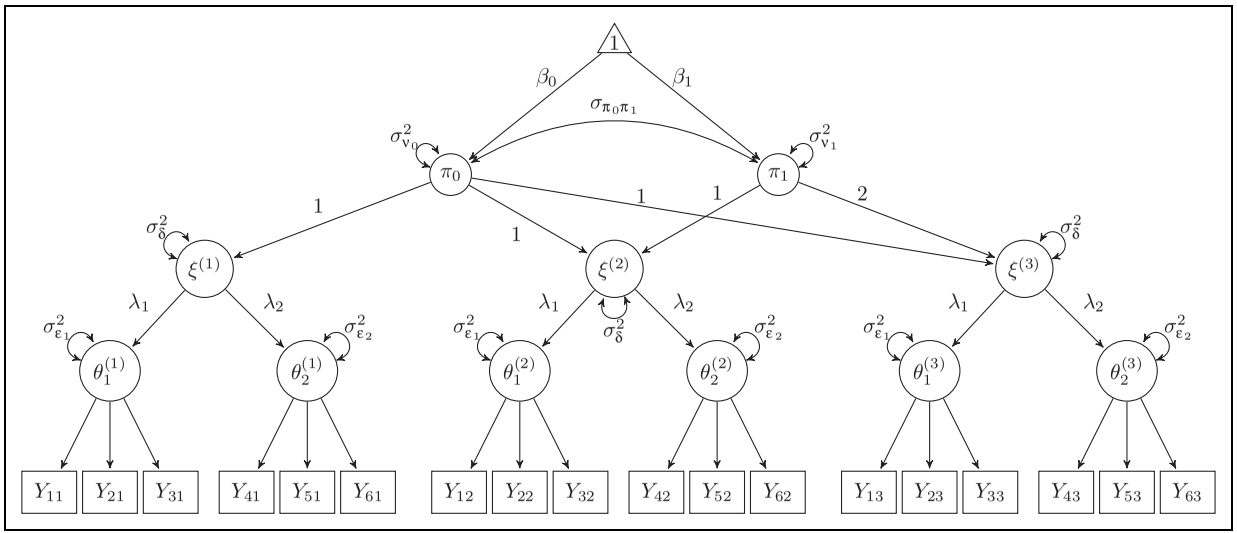

FIGURE 3. *A path diagram for the longitudinal, higher order item response theory model with three items per domain-level trait, two domain-level traits, and three time points.* π_0 *and* π_1 *represent the random intercept and random linear slope parameters per person.* β_0 *and* β_1 *are the population means of* π_0 *and* π_1*, respectively.*

intercept and slope parameters to take the predetermined structure of $\lambda_k \pi_{i0}$ and $\lambda_k \pi_{i1}$ due to the functional form of the model.

Assuming either the same sets of items are repeatedly administered or that the test includes shared items between adjacent time points for all domains, the minimum model identifiability constraints include:

1. All of the residuals having mean 0 (i.e., $E(\delta_i^t) = 0$ for all $t = 1, \ldots, T$).
2. The mean of the person-specific intercept parameters being set to 0 (i.e., $\mu_{\pi_{i0}} = \beta_0$). The purpose of this constraint is to specify the location of ξ^t at $t = 1$.
3. All of the residuals in the measurement model having mean 0 (i.e., $E(\epsilon_{ik}^t) = 0$ for all $k = 1, \ldots, K$ and $t = 1, \ldots, T$) in Equation 9. This assumption is typical for a factor regression model and made in de la Torre and Song (2009).
4. The residual variances at the first time point being set to a constant (i.e., $\sigma^2_{\epsilon_{ik}^1} = c_k^1$, where c_k^1 is a user-specified constant). This constraint is necessary to establish the scale of the θs in the model. Justification for this constraint is similar to justification for the similar constraint in the L-UIRT and L-MIRT models and is due to $\boldsymbol{\theta}_i^t$ being endogenous to the model, so that its variance can only be fixed indirectly by setting its residual variance. Only the variance at a single time point needs to be fixed, as the variance of $\boldsymbol{\theta}_i^t$ at the remaining time points are determined via the linking items.
5. One of the loading parameters, λ_k for some some k ($k = 1, \ldots, K$) being set to a constant, assuming that λ_k is invariant over time. The remaining $(K - 1)$ loading parameters are freely estimable.

The first two constraints are essentially the same as the first two constraints for both the L-UIRT model and the L-MIRT model described earlier. The remaining constraints are unique to the L-HO-IRT model. The last constraint is similar to the "reference indicator" constraint in factor analysis. That is, the variance of a factor can be determined by fixing the loading of one marker indicator. Here, the "marker indicator" is one of the first-order factors, θ_{ik} for some k ($k = 1, \ldots, K$), and the "factor" is ξ_i. Readers of de la Torre and Hong (2010) may notice that they imposed a different constraint for the same purpose, namely

$$\text{var}(\epsilon_{ik}^t) = 1 - \lambda_k^2. \tag{12}$$

They argued that the variance of θ_{ik} is typically assumed to be 1, and the assumption from Equation 12 results in a variance of ξ_i also assumed to be 1. Thus, by way of this constraint, both the first-order and second-order factors would be on the same scale. The motivation of de la Torre and Hong (2010) is not relevant to our current discussion, as the variance of θ_{ik} is not assumed to be a constant over time (and might have good reason not to be given the type of change observed). If requiring standardized loading parameters, one could calculate a simple linear transformation of λ_k, that is $\lambda_k^\star = \lambda_k \times \frac{\sigma_{\xi^t}}{\sigma_{\theta_k^t}}$. Moreover, in M*plus*, the equality constraints in Equation 12 can only be specified with

maximum likelihood estimation (MLE) but not with the Bayesian estimation option. Note that when anchor items are precalibrated with known parameters, then only the first, third, and last constraints are necessary to identify the model.

2.4. Applications of the Models

Applying one of the above models versus another depends mostly on the hypothesized factor structure of the latent traits. Higher-order models are often applicable in contexts where a measurement instrument assesses several related constructs that can be accounted for by one or more underlying second-order factors (Chen et al., 2006). For instance, a common scale to measure "quality of life" is composed of four *subscales* that each presume to measure a distinct first-order factor: mental health, cognition, vitality, and health worry (Chen et al., 2006). The covariance between each pair of first-order factors can be explained by a higher order factor, which is usually called "global quality of life." Similarly, educational measures are often constructed to assess several, separate but correlated, content domains that can be partially explained by a more general ability. For instance, a mathematics test may have items measuring numerical computation skills and data analysis skills (Reckase, 2009, p. 232). Both of these are examples of content-based multidimensionality rather than strict construct-based multidimensionality.

In practice, one cannot typically distinguish between content multidimensionality and construct multidimensionality because content-based subscales often measure distinct constructs. Yet certain content-based domains sometimes have exceedingly high correlations, implying that these domains essentially measure the same skill or construct (Reckase, 2009). In cases like these, one should always provide evidence that combining domains makes substantive sense or yields a better fit than keeping those domains separate.

Although a correlated-factor MIRT model will always fit data generated from the HO-IRT model, the higher order model has at least four advantages for being preferred in practice: As compared with the correlated-factor MIRT model, the HO-IRT model (1) parsimoniously explains the covariance between lower order factors (Gustafsson & Balke, 1993; Rindskopf & Rose, 1988), (2) separates the variance in the lower order factors shared by the common higher order factor from the unique variance of the lower order factors, (3) simplifies model estimation due to the exploitation of the dimension reduction technique (as described in the next section), and (4) allows for potential construct shifts over time.

To elaborate on the last point, assume teachers want to track students' ability in a general subject area such as math knowledge. If math knowledge is a unidimensional trait, it can be measured directly by a set of items, and if the teacher is not interested in measuring any specific subareas of mathematics, then the L-UIRT model is sufficient. However, math knowledge might relate to a number of specific content areas that teachers might also wish to track. For example, Table 1

TABLE 1.
Mathematics Common Core Domains by Grade (K–4)

	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4
Domain 1	Counting and cardinality				
Domain 2	Operations and algebraic thinking	Operations and algebraic thinking	Operations and algebraic thinking	Operations and algebraic thinking	Operations and algebraic thinking
Domain 3	Number and operations in base 10	Number and operations in base 10	Number and operations in base 10	Number and operations in base 10	Number and operations in base 10
Domain 4	Measurement and data	Measurement and data	Measurement and data	Measurement and data	Measurement and data
Domain 5	Geometry	Geometry	Geometry	Geometry	Geometry
Domain 6				Number and operations—Fractions	Number and operations—Fractions

presents the content coverage of the mathematics common core domains across five domains. The domains (such as Domain 5 "Geometry" and Domain 4 "Measurement and Data") are expected to be taught and developed in every grade from Kindergarten–4. Student growth in these domains can be tracked across all five grades. However, the required content coverage shifts from grade to grade, and many domains only appear in limited grades. For instance, Domain 1 ("Counting and Cardinality") is expected to be assessed only in Kindergarden, whereas Domain 6 ("Numbers and Operations-Fractions") does not emerge until Grade 3. In these cases, the L-MIRT model and L-UIRT model overlook crucial details. In particular, the L-MIRT model (Hsieh et al., 2010) essentially assumes a constant set of traits measured over time. For this relatively straightforward example, the domains are designed to change over time.

However, when indeed the same sets of domains are measured overtime, the L-MIRT model is preferred because the L-HO-IRT model is parametrically more restricted than the L-MIRT model. That is, any growth patterns in the lower level traits that can be captured with the L-HO-IRT model can ultimately be captured with the L-MIRT model. Yet, if the multidimensional (lower level) constructs each change differently over time, then the L-HO-IRT model would no longer fit the data, and one should use the L-MIRT model. For instance, if certain domain-level traits grow linearly, whereas others grow in a piecewise fashion, then one should no longer use the L-HO-IRT model due to the restrictions implicit in Equation 10. On the other hand, the L-MIRT model can handle different growth patterns if needed.

When assessing change over time, one must consider whether the measures retain measurement invariance. Often, practitioners use the exact same scale on

multiple occasions. This practice can ensure that identical constructs are continuously assessed and that the metric of measurement remains the same over time. However, out of necessity, scales often differ across repeated measurements due to the need for "developmentally appropriate measures" (Widaman, Ferrer, & Conger, 2010). Adjusting the scale to consider the typical range of traits over repeated measurements can help avoid ceiling and floor effects.

Determining whether the same construct, measured by multiple indicators, has the same meaning and metric over time falls under the rubric of measurement invariance (Widaman et al., 2010), and is often referred to, especially in a longitudinal setting, as longitudinal invariance. The factorial invariance of longitudinal measures is paramount in evaluating the change in behavior over time (McArdle, 2001; McArdle & Hamagami, 2001; Meredith & Tisak, 1990; Widaman & Reise, 1997). Using the same set of items or a set of anchor items (Grimm, Kuhl, & Zhang, 2013) partially satisfies longitudinal invariance. A thorough examination of longitudinal invariance is beyond the scope of this article. Interested readers can refer to Teresi (2006), Isiordia and Ferrer (2018), Liu et al. (2017) for details regarding invariance assumptions of L-UIRT, L-MIRT (i.e., CUFFS), and L-HO-IRT, respectively.

3. Model Estimation

Within the general framework of SEM, the L-IRT models can be viewed as a multilevel LGC model with the lowest level represented by categorical indicators. Unsurprisingly, the L-IRT models can also be motivated from the framework of generalized linear models (McCullagh & Nelder, 1989), a conceptualization favored within biostatistics. The most common methods for estimating multilevel models are based on integrating the likelihood over the distribution of random effects, which is often referred to as marginal likelihood estimation. For instance, in the L-HO-IRT model, the overall- and domain-specific latent abilities as well as the latent intercepts and slopes represent the random effects over which to integrate. Because analytical integrals often do not exist for these types of models, researchers frequently adopt one of the two classes of methods. One could either approximate the integrand analytically or evaluate the integral via numerical approximation. The first approach includes Laplace's method of linearizing the integrand via a sixth-order Taylor series approximation (called "Laplace 6") as well as quasi-likelihood methods such as marginal quasi likelihood (MQL; Goldstein, 1991; Goldstein & Rasbasch, 1996) and penalized quasi likelihood (PQL; Breslow & Clayton, 1993; Laird, 1978). Because the performance of PQL and MQL depends on the validity of a normal approximation, these methods tend to perform poorly when the observed data are markedly nonnormal (Rodriguez & Goldman, 1995; Tuerlinckx, Rijmen, Verbeke, & Paul De Boeck, 2006) and are thus typically not recommended for use in IRT models with binary responses. The second approach includes ML

using Gauss–Hermite quadrature, adaptive quadrature, and simulation methods such as the Monte Carlo expectation-maximization (EM) algorithm (Wang & Xu, 2015).

However, ML estimation via the EM algorithm is known to converge slowly in many applications (e.g., Meng & van Dyk, 1997) and is computationally intensive when the number of latent variables is large. Bayesian estimation using Markov chain Monte Carlo (MCMC) with diffuse (or noninformative) priors (Patz & Junker, 1999) is an alternative to EM (Huang, 2013; Wang & Nydick, 2015) and is usually preferred for complex models.

All of the above estimation methods are based on full information, in that the likelihood is constructed directly from the raw response pattern. Alternatively, one could adopt limited information estimation methods, such as modified weighted least squares (WLS) estimation. Rather than basing the likelihood on the complete response pattern, modified WLS estimates model parameters via the first four moments of the response contingency table. By avoiding the time-consuming numerical integration or sampling steps of the full information methods, WLS leads to much faster convergence. However, WLS is known to yield inaccurate estimation with small sample sizes or large amounts of missing data (e.g., Forero & Maydeu-Olivares, 2009). Moreover, the parameter estimates from WLS are not as efficient as a full information method (Muthén & Asparouhov, 2015). Given these limitations, WLS is not discussed further in this article.

In the following subsections, we describe estimating the L-IRT models in M*plus* with ML or MCMC methods. M*plus* software was chosen due to being widely used in social science research. Other IRT estimation software packages, such as *flex*MIRT (see Paek et al., 2016, for details on how to estimate similar models to those described in this article), or general-purpose estimation packages, such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), should also be able to recover L-IRT-based model parameters. Interested readers could refer to Curtis (2010) or Isiordia and Ferrer (2018), which present BUGS code and R code (using the "lavaan" package, see Rosseel, 2012), respectively, for estimating a subset of L-IRT models. Details of estimating L-IRT models using WLS are explained in Wang, Kohli, and Henn (2016).

3.1. MLE

When using M*plus*, one must specify the model estimation method in the `ANALYSIS` section of the input script. If estimating IRT-based item parameters with MLE, include the following `ANALYSIS` statement:

```
ANALYSIS: TYPE = GENERAL;
          ESTIMATOR = MLR;
          LINK = LOGIT;
          INTEGRATION = MONTECARLO;
```

TABLE 2.
*Number of Continuous Dimensions and Dimensions of Numerical Integration for Different Models and Methods (*T *denotes the number of time points,* K *denotes the number of lower-order latent traits,* q *denotes the number of random effects)*

Models	Number of Continuous Latent Variables	Dimensions of Numerical Integration (M*plus* Default)	Dimensions of Numerical Integration (Analytic Dimension Reduction)
L-UIRT	$T + q$ (6)	T (4)	$q + 1$ (3)
L-MIRT	$T \times K + q \times K$ (30)	$T \times K$ (20)	$q \times K + 1$ (11)
L-HO-IRT	$T \times K + T + q$ (26)	$T \times K$ (20)	$q + 2$ (4)

Note. IRT = item response theory; L-UIRT = longitudinal unidimensional IRT; L-MIRT = longitudinal multidimensional IRT; L-HO-IRT = longitudinal higher order IRT.

As indicated in the last line of the previous statement, we recommend using M*plus*'s `MONTECARLO` integration routine for the numeric integration. Without including the `INTEGRATION` line, M*plus* would default to use rectangular (trapezoid) numerical integration with either 15 adaptive quadrature points per dimension, or 30 to 50 nonadaptive quadrature points per dimension (Chapter 14, M*plus* User Guide). Although adaptive numeric integration is computationally faster, if the data have outliers or nonnormally distributed latent traits, it may yield unstable results. If estimating a model with one to three dimensions of integration, the default quadrature-based numerical integration algorithm usually results in precise estimates. Conversely, `MONTECARLO` integration does not yield as accurate estimates of parameters for low dimensions of integration but is much more efficient for higher dimensional integration.

Table 2 illustrates the dimensions of numeric integration for each of the three models with values in parentheses assuming that $T = 4$, $K = 5$, and $q = 2$. As shown in Table 2, the number of continuous latent variable per model (the second column in Table 2) is simply the number of latent factors (including the first-order and second-order latent traits) plus the number of random effects (the person-specific intercepts and slopes). The dimensions of integration (the third column in Table 2) include only those factors that have categorical indicators (the θs) as opposed to higher level factors (the ξs) or random effects. According to the M*plus* User Guide (p. 527), closed form solutions may exist for integrating out latent factors with continuous indicators, such as the ξs or random effects, so that the numerical integration approximation is no longer needed. Nonetheless, the number of dimensions of integration for all three longitudinal models is prohibitively large.

The right-most column in Table 2 indicates the dimensions of integration if using an analytic dimension reduction technique. Analytic dimension reduction is often used to rearrange terms in the marginal likelihood integral to yield a series of integrals, each of much lower dimension than the original integral (Cai, Yang, & Hansen, 2011; Gibbons & Hedeker, 1992; Rijmen, Vansteelandt, & de Boeck, 2008). Applying a dimension reduction technique to the L-UIRT model, rewrite Equation (3) as

$$\frac{1}{1+\exp\left[-a_j^t\left(\theta_i^t-b_j^t\right)\right]}=\frac{1}{1+\exp\left[-a_j^t\left(\pi_{0i}+\pi_{1i}\times(t-1)+\delta_i^t-b_j^t\right)\right]}. \tag{13}$$

If assuming that δ_i^ts are uncorrelated across pairs of time points, then one need only integrate out π_{0i}, π_{1i}, and δ_i^t, resulting in a three-dimensional integral, for a given item (Paek et al., 2016). The same arguments lead to a similar dimension reduction solution to the L-MIRT model. The results for the L-MIRT model in Table 2 are based on the assumption that the residual covariance matrix of $\boldsymbol{\delta}_i$ is a diagonal matrix.[1]

The L-HO-IRT model has a different dimension reduction solution given the addition of the higher level trait. First, write the HO-IRT IRF as

$$\frac{1}{1+\exp\left[-a_{j_k}^t\left(\theta_{ik}^t-b_{j_k}^t\right)\right]}=\frac{1}{1+\exp\left\{-a_{j_k}^t\left[\lambda_k\left(\pi_{0i}+\pi_{1i}\times(t-1)+\delta_i^t\right)+\epsilon_{ik}^t-b_{j_k}^t\right]\right\}}, \tag{14}$$

where a_{j_k} and b_{j_k} denote item parameters for item j measuring domain k. In Equation 14, the only additional random effect to integrate out of the likelihood equation is ϵ_{ik}^t. Because all ϵ_{ik}^ts are assumed uncorrelated across time, then generalized dimension reduction yields a four-dimensional integral (π_{0i}, π_{1i}, and δ_i^t as before, as well as ϵ_{ik}^t). Note that this dimension reduction technique can only be applied if the residuals from the growth curve model, δ_i^t, are uncorrelated across time. If estimating models with correlated residuals (such as an autoregressive model), this dimension reduction technique can no longer be applied.

Advantages of estimating parameters using the EM algorithm, as compared with Bayesian methods, in M*plus* include: (1) being able to estimate the three-parameter logistic (3PL) model rather than only being able to estimate one or two parameter normal ogive models, (2) providing comparative model fit indices such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), and (3) being able to impose equality constraints on model parameters. Note that these limitations of Bayesian methods are not necessarily inherent to the methods themselves, only to the application of those methods in M*plus*. Due to the high-dimensional integration, we have had more success estimating the L-IRT models with the MCMC option in M*plus*. Researchers and practitioners

should always keep in mind complexity and feasibility when choosing a model and corresponding estimation algorithm.

3.2. MCMC

If estimating IRT-based item parameters with MCMC, include the following ANALYSIS statement:

```
ANALYSIS: ESTIMATOR = BAYES;
          CHAINS = 1;
          FBITER = 50000;
          POINT = MEAN;
```

In the above statement, the FBITER line denotes the fixed number of iterations for each Markov chain (i.e., the chain length). If FBITER is not specified, the chain will stop once convergence is reached with the default convergence criterion being a potential scale reduction (PSR; Gelman & Rubin, 1992) at or below 1.05 (see M*plus* User Guide, 1998-2011, p. 640). After 50,000 iterations, POINT = MEAN indicates that the posterior mean will be used as the point estimate of the model parameters.

The next section provides a real data example of applying M*plus* (Version 8 used in this study) to estimate parameters of data that fit the L-IRT model. A corresponding simulation study, demonstrating parameter recovery of the three L-IRT models, is included as an Appendix in the online version of the journal to this article.

4. A Real Data Example

The current section applies the three L-IRT models to a real data example. The purpose of this demonstration is to illustrate the potential application of each model as well as the information each model provides to researchers and practitioners. For this purpose, we adopted and analyzed a series of math assessments that students in one Midwest state took between 2009 and 2012. These students were assessed in each of Grades 3 through 6 using a five-dimensional, simple-structure test with precalibrated item parameters. The five dimensions had been termed "number and operation," "geometry and spatial sense," "data analysis, statistics, probability," "measurement," and "algebra, functions, and patterns," respectively. Students took 57 items in 2009 (with 23, 9, 7, 11, and 7 items, respectively, measuring each dimension) and 52 items in each of the three subsequent years (with 23, 9, 7, 11, and 7 items, respectively, measuring each dimension). After initial data cleaning, only $N = 327$ students had a complete set of mixed responses (i.e., including both correct and incorrect responses) for sets of items on each dimension at every time point.[2]

Due to different sets of items being administered in each year, common-item linking is not possible. However, precalibrated anchor items were embedded within each of the five dimensions across all 4 years and are all on the same scale. Because of fixing known anchor items, many of the identifiability constraints need not be explicitly specified (see the model description section for additional details). Only λ_1 in the L-HO-IRT model must still be specified, and we set $\lambda_1 = 1$ to fix the scale of ξ. All growth models were assumed to have only random intercepts and slopes (see the spaghetti plots below for linearity of time on θ and ξ). Moreover, all responses were assumed to conform to the 2PL IRT model.[3] For estimation, an MCMC algorithm was run in M*plus* with a Markov chain length (FBITER) fixed to 30,000 with the first half of the iterations discarded as burn-in by default. In all cases, the PSR for all model parameters were below 1.03, implying successful chain convergence.

To evaluate global model fit in Bayesian models with categorical outcome variables, M*plus* provides the Bayesian posterior predictive *p* value (Kaplan & Depaoli, 2012; Muthén, 2010). In our case, the Bayesian *p* value for the L-UIRT, L-MIRT, and L-HO-IRT[4] models were estimated to be .103, .081, and .106, respectively, implying that all three models yielded acceptable global fit. Note that other Bayesian software packages such as JAGS (Plummer, 2003) provides the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linden, 2002) for model comparison. M*plus* does not yet include DIC for models with categorical indicators.

Table 3 presents the parameter estimates from the three L-IRT models. Because of fixing $\lambda_1 = 1$ in the L-HO-IRT model, parameter estimates from this model may not be on the same scale as those from the L-UIRT and L-MIRT models. Even though parameter estimates are not strictly comparable across models, we can still make some general statements based on Table 3: (1) The fixed effect of time is positive, implying an increase in average ability over time; (2) the intercept and slope combined variances (i.e., $\mathbf{Z}\Sigma_{\nu}\mathbf{Z}^{\mathrm{T}}$, where $\mathbf{Z}$ is the design matrix defined in Equation 1, indicating the dependent variable variance explained by random effects) greatly exceed the residual variance in the growth part of the model (i.e., $\sigma^2_{\delta_i}$), which evidences the linear functional form being sufficient to capture the latent growth pattern; (3) the random intercepts vary more than the random slopes, and there is a moderate negative correlation (of $-.3$ to $-.4$) between random intercepts and random slopes, implying larger differences in initial ability than in growth rates. This moderate negative correlation between initial state and growth is interesting and implies that the gap between high- and low-performing students decreases over time. Even though one cannot directly compare parameter estimates from the L-MIRT and L-HO-IRT models, the intercept variances being larger for Domains 2 and 3 in the L-MIRT model (i.e., .255 and .267 in Table 3) are consistent with the λs being relatively lower for these two domains (i.e., .768 and .734) in the L-HO-IRT

TABLE 3.
Structural Model Parameter Estimates for Three Different Models

Models	NP	Fixed Effects $(\beta_0 \quad \beta_1)$	Random Effects $\begin{pmatrix} \sigma^2_{\pi_{0i}} & \\ \sigma_{\pi_{0i}\pi_{1i}} & \sigma^2_{\pi_{1i}} \end{pmatrix}$	Others
L-UIRT	275	$(-.653 \quad .472)$	$\begin{pmatrix} .081 & \\ -.008 & .005 \end{pmatrix}$	$\boldsymbol{\sigma}^2_{\delta_i} = (.048 \quad .063 \quad .046 \quad .015)$
L-MIRT	351	$\begin{pmatrix} -.652 & .509 \\ -.633 & .428 \\ -.421 & .356 \\ -.659 & .444 \\ -.795 & .524 \end{pmatrix}$	$\begin{pmatrix} .145 & .014 \\ .255 & .039 \\ .267 & .038 \\ .169 & .042 \\ .120 & .017 \end{pmatrix}$	$\boldsymbol{\sigma}^2_{\delta_i} = \begin{pmatrix} .051 & .087 & .052 & .019 \\ .063 & .066 & .025 & .037 \\ .032 & .048 & .009 & .029 \\ .035 & .055 & .042 & .031 \\ .054 & .010 & .031 & .013 \end{pmatrix}$
L-HO-IRT	299	$(-.702 \quad .514)$	$\begin{pmatrix} .102 & \\ -.009 & .007 \end{pmatrix}$	$\boldsymbol{\lambda} = \begin{pmatrix} 1^* \\ .768 \\ .734 \\ .882 \\ .935 \end{pmatrix}$
				$\boldsymbol{\sigma}^2_{\delta_i} = (.052 \quad .071 \quad .061 \quad .015)$
				$\boldsymbol{\sigma}^2_{\epsilon_i} = \begin{pmatrix} .028 & .031 & .016 & .017 \\ .121 & .091 & .088 & .044 \\ .018 & .031 & .013 & .011 \\ .059 & .020 & .018 & .034 \\ .059 & .012 & .013 & .008 \end{pmatrix}$

Note. NP denotes the number of free parameters in each model. The covariances between random intercepts and random slopes from the L-MIRT model are omitted to save space because they are between −.01 and .01. "*" denotes a fixed constant. IRT = item response theory; L-UIRT = longitudinal unidimensional IRT; L-MIRT = longitudinal multidimensional IRT; L-HO-IRT = longitudinal higher order IRT.

model. Thus, estimation patterns persist regardless of lack of direct comparability of parameter magnitudes.

In contrast to the L-HO-IRT model, the L-UIRT and L-MIRT models can be directly compared in this case due to anchor items setting the scale for the lower order traits. From Table 3, one can see that averaging the intercepts and slopes from the L-MIRT model leads to estimates similar to those from the L-UIRT model. Yet the variance of the intercept and slopes from the L-MIRT model is much larger, implying that evaluating individual performance at the domain level leads to higher variability than assuming that responses are all generated from a single, common trait. That said, if a test is constructed across several domains, considering domain-level growth patterns may reveal subgroup differences

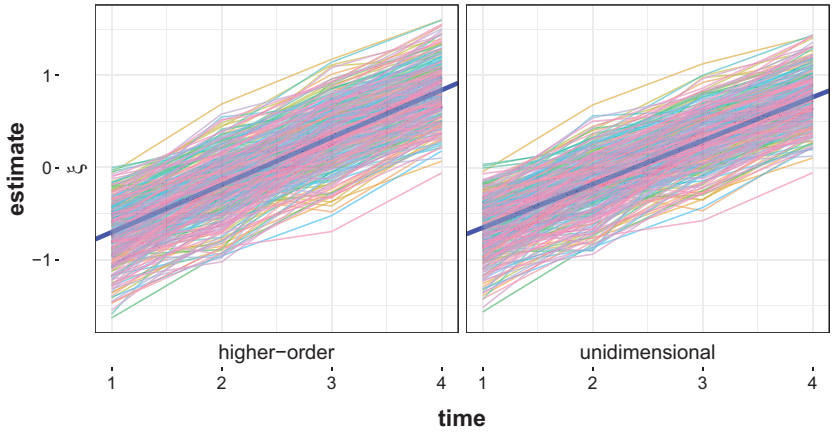

FIGURE 4. *A spaghetti plot, illustrating the linear trend of ξ (overall-level ability) on math between Grades 3 and 6 for* N $= 327$ *students. The left panel is obtained from the longitudinal higher order item response theory model, and the right panel is from the longitudinal unidimensional item response theory model. The bolded, slanted line in the center of the spaghetti depicts the estimated fixed effect of time.*

otherwise diminished if assuming responses came entirely from a unidimensional trait.

Figure 4 presents a spaghetti plot of the overall ability across time for $N = 327$ students using the L-HO-IRT model (left) and the L-UIRT model (right). Unsurprisingly, the lines in the right panel are slightly closer together than the lines in the left panel, which is consistent with the results in Table 3 that the variance of the random slopes is slightly higher from the L-HO-IRT model. Figure 5 presents the spaghetti plot of the domain-specific abilities across time using the L-HO-IRT model (upper) and the L-MIRT model (lower). As shown in Figure 5, aside from minor differences, the overall growth lines and the individual growth trajectories from both models exhibit similar patterns. One anomaly worth mentioning is that the individual growth curves from the L-HO-IRT model tend to fluctuate quite a bit more than the growth curves from the L-MIRT model. The L-MIRT model growth curves (for all but $k = 1$ and $k = 4$) tend to follow strict lines. This result is due to where the growth trajectory is imposed. With respect to the L-HO-IRT model, the growth trajectory is fit to the θs only indirectly (due to the θs relationship with ξ) as reflected in Equation 11. Because of this indirect effect, the residual variance of θ $\left(\sigma^2_{\upsilon^t_{ik}} = \lambda^2_k \sigma^2_{\delta^t_i} + 1 - \sigma^2_{\epsilon^t_{ik}}\right)$ could be large, and the individual growth trajectories might exhibit some departure from a strict line. Conversely, with respect to the L-MIRT model, a growth line is imposed directly on the individual θs (see Equation 5). Due to a small estimated

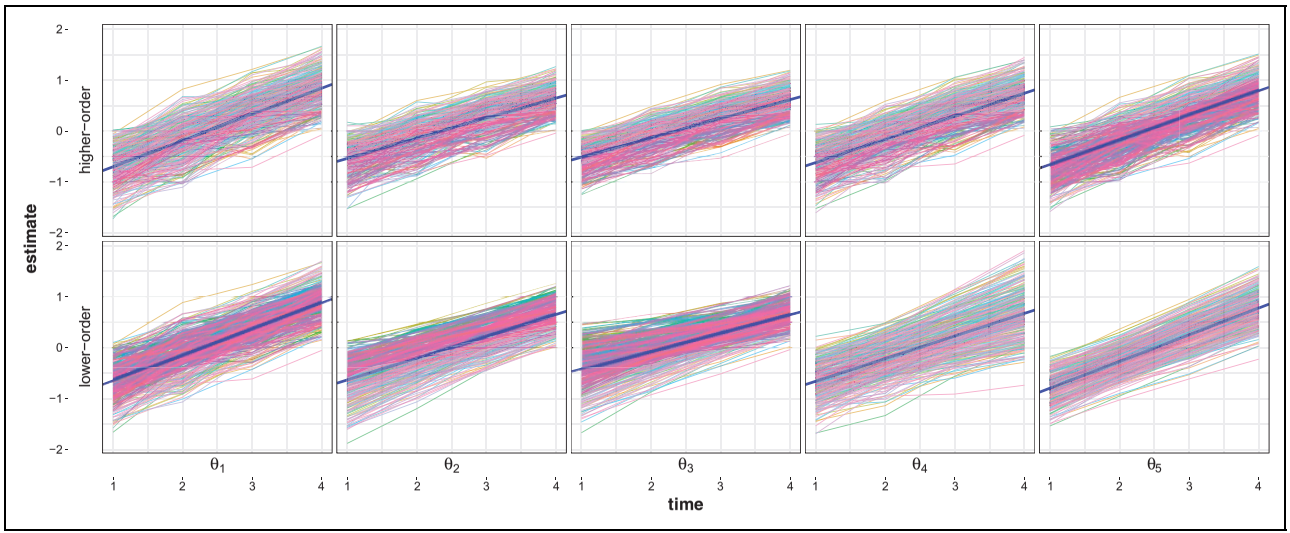

FIGURE 5. *A spaghetti plot, illustrating the linear trend of* θ *(domain-level ability) on math between Grades 3 and 6 for* $\mathrm{N} = 327$ *students. The upper panels are obtained from the longitudinal higher order item response theory model, and the lower panels are from the longitudinal multidimensional item response theory model. The bolded, slanted line in the center of the spaghetti depicts the estimated fixed effect of time.*

residual variance in the θs (between .007 and .088), the domain abilities were estimated to be close to the trajectory line. Note that the figures reinforce linearity in the average growth pattern over time, which was implied earlier by comparing the slope/intercept variances to the residual variances from Table 3.

5. Conclusion

Many teachers, administrators, and policymakers require the measurement of student growth. Teachers can use estimated growth to modify lesson plans based on strategies of improvements. Administrators can use estimated growth to examine school performance and help make budgetary decisions. In either case, one must ensure estimates are accurate across several, possibly correlated, ability dimensions. Several L-IRT models haven been proposed for different purposes. These L-IRT models all share the same form and contain two components: (1) an IRT measurement model for each measurement occasion and (2) a LGC model imposed on the latent trait, quantifying the intraindividual developmental trajectories. In this article, we reviewed three specific types of L-IRT models with the goal of demonstrating appropriate applications of these models for longitudinal assessment. We also illustrated fitting different models with a commonly used software package.

Among the three models, the L-UIRT model is the simplest and has been the most extensively studied in the literature (e.g., Andersen, 1985; Embreston, 1991, Grimm et al., 2013; McArdle et al., 2009; von Davier et al., 2011; Wang et al., 2016; Wilson et al., 2012). In contrast to the L-UIRT model, which tracks change in a unidimensional latent trait, the L-MIRT model describes change in multiple, correlated latent traits (see Paek et al., 2016). Compared to models that directly model change in the lower level abilities, the L-HO-IRT model includes two unique features. First, because the HO-IRT model captures the hierarchical nature of learning, the L-HO-IRT model simultaneously models the growth trajectories of both overall- and domain-specific abilities. Second, as described earlier in this article, the L-HO-IRT model allows for a shift in domain coverage over time, as long as one carefully verifies the second-order longitudinal invariance requirement (e.g., Chen et al., 2006; Liu et al., 2017). Allowing for a shift in the domain coverage over time is extremely important in educational measures, as one typically finds more advanced domains added and basic domains eliminated as students complete more schooling. Furthermore, a higher order model allows one to find trends at the individual, domain level. Domain-level information can hint at particular academic subjects that improve the most over particular grades. For instance, in our real data example, θ_1 and θ_5 tended to improve the most over time, and θ_3 tended to improve the least (assuming, of course, that the location and scale across dimensions are comparable). With a L-HO-IRT model, one can obtain estimates of overall trends as well as delve into individual dimensions underlying complex assessments.

In terms of model estimation, we provided a thorough discussion of the analytical dimension reduction techniques that are available to alleviate high-dimensional integration challenges of marginal MLE (MMLE). Even after dimension reduction, the number of integration dimensions can still be high. In this case, the Metropolis–Hastings Robbins Monro algorithm (Cai, 2010) or the MCMC algorithm can be used in lieu of MMLE via EM. Given that the L-MIRT and L-HO-IRT are less studied in the literature, a simulation study was conducted to provide a thorough quality control check on the precision in estimating model parameters (refer to the Supplementary File in the online version of the journal for details of the simulation, which evaluated the recovery of both structural parameters and individual latent traits/growth parameters). When examining simulation results, all model parameters were adequately recovered, and the generating model evidenced adequate model fit. Even with the supporting evidence from the simulation study, interested users of the L-HO-IRT and L-MIRT models should keep in mind that both of these models should only be applied when there are sufficient items per domain, otherwise the domain-level θs and the resulting higher order factors (i.e., ξ and growth parameters) would not be reliably estimated.

This article serves two purposes. First, no prior paper has explicitly documented and reviewed the three popular L-IRT models as well as their identifiability constraints with and without known item parameters. Including this information has profound didactic value for practitioners who wish to apply the models to their own data. Sample M*plus* code is provided in the Appendix in the online version of the journal for each model for readers' reference. Second, this article is the first attempt to thoroughly compare and demonstrate the applicability of each of the discussed models. Even though these models can adequately capture changes in typical longitudinal measures, they are by no means exhaustive. A handful of other longitudinal models exist, such as the two-tier model (Cai, 2010), in which nuisance factors are introduced to account for residual dependencies between common items over time, or the item-level growth curve model (Paek et al., 2016), in which growth rates for different items can differ and therefore be described and examined.

Regardless of chosen model, constructing and estimating growth using L-IRT can improve the measurement of educational outcomes and thus provide educators with tools they need to better help students learn. Currently available software packages can estimate growth across a wide variety of measurement models (e.g., 1PL, 2PL, 3PL, unidimensional, multidimensional, and higher order) and LGC models (i.e., Equations 1 and 4). Interested practitioners should be cognizant of the different estimation methods offered in each of the programs and to choose the method appropriate for the problem at hand, especially given complex models with many estimable parameters. For instance, the discussed analytic dimension reduction technique is only relevant to MML estimation approaches but not to the Bayesian MCMC estimation approach commonly used to estimate

parameters of complex models. Software packages such as M*plus* may not automatically use a given dimension reduction unless the command file (or source script) is written with dimension reduction in mind.[5] Hence, understanding the logic of dimension reduction can help with constructing the command file or script processed by the algorithm and greatly reduce computation time.

Although this didactic offers sufficient technical details for three popular L-IRT models for researchers and practitioners to use those models in their own research, two relevant topics were outside the scope of the current discussion. First, LCG models with intrinsically nonlinear growth patterns were not discussed because this family of models is not currently included in a majority of software packages for LCG model estimation. An example of this kind of model is a "piece-wise growth curve model with unknown knots" (e.g., Kohli, Hughes, Wang, Zopluoglu, & Davison, 2015). Second, we have not discussed how to evaluate global model fit. Although most SEM software packages will output one or multiple absolute fit indices, few studies have examined appropriate cutoffs for these indices in determining adequate fit. Moreover, the DIC that is often used with MCMC can take different forms. The first-level conditional DIC provided by WinBUGS may not always provide the best estimates of model fit, whereas a second-level joint DIC might be more appropriate for multilevel IRT models (Zhang, Tao, & Wang, 2019). A thorough examination of model fit for L-IRT models is needed to ensure credible conclusions drawn from any model-based results.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170042 (or R305D160010) awarded to the University of Washington and the Spencer/National Academy of Education Post-doc Fellowship (2014).

Notes

1. If, on the other hand, the residual covariance matrix of $\boldsymbol{\delta}_i$ is a block diagonal matrix, allowing the residuals from different latent traits to correlate at a given time point, then the dimensions of numerical integration would be $(q + 1) \times K$.

2. The complete data set is available for download on www.placeholder.com.
3. M*plus* can estimate three-parameter logistic model parameters using only the marginal maximum likelihood estimation/EM algorithm, which becomes exceedingly slow when the number of integration dimensions is large, such as in the longitudinal multidimensional item response theory or longitudinal higher order item response theory models considered in this article.
4. Originally, we ran the model allowing λ_2 to λ_K to differ across time. Relaxing the invariance assumption resulted in a posterior predictive *p* value changed by .001. Because imposing an invariance assumption still yields a *p* value > .05, we decided to base our results and discussion on the invariance model.
5. Please see an example for the higher order item response theory model at www.placeholder.com.

References

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

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis*, *95*, 1–22.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley-Interscience.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9–25.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221–248.

Chen, F., Sousa, K., & West, S. (2006). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471–492.

Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software*. Retrieved from https://www.jstatsoft.org/article/view/v036c01

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT model approach. *Applied Psychological Measurement*, *34*, 267–285.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*, 620–639.

DeYoung, C. (2006). Higher-order factors of the Big Five in a multi-information sample. *Journal of Personality and Social Psychology*, *91*, 1138–1151.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (Eds.). (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Edwards, M., & Wirth, R. (2009). Measurement and study of change. *Research in Human Development*, *6*, 74–96.

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

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*, 275–299.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Golay, P., & Lecerf, T. (2011). Orthogonal higher order structure and confirmatory factor analysis of the French Wechsler Adult Intelligence Scale (WAIS-III). *Psychological Assessment*, 23, 143–52.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, *78*, 45–51.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, *159*, 505–513.

Grimm, K. J., Kuhl, A. P., & Zhang, Z. (2013). Measurement models, estimation, and the study of change. *Structural Equation Modeling*, *20*, 504–517.

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*, 407–434.

Hsieh, C.-A., von Eye, A. A., & Maier, K. S. (2010). Using a multivariate multilevel polytomous item response theory model to study parallel processes of change: The dynamic association between adolescents' social isolation and engagement with delinquent peers in the national youth survey. *Multivariate Behavioral Research*, *45*, 508–552.

Huang, H. Y. (2013). *Measuring latent growth under the multilevel higher-order item response theory model*. Paper presented at 2013 Annual Meeting of National Council on Measurement in Education, San Francisco, CA.

Huang, H. Y. (2015). A multilevel higher order item response theory model for measuring latent growth in longitudinal data. *Applied Psychological Measurement*, *39*, 362–372.

Huang, H. Y., & Wang, W. C. (2014). Multilevel higher-order item response theory models. *Educational and Psychological Measurement*, *74*, 495–515.

Isiordia, M., & Ferrer, E. (2018). Curve of factors model: A latent growth modeling approach for education research. *Educational and Psychological Measurement*, *78*, 203–231.

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Eds.), *Handbook of structural equation modeling* (pp. 650–673). New York, NY: Guilford Press.

Kohli, N., Hughes, J., Wang, C., Zopluoglu, C., & Davison, M. L. (2015). Fitting a linear-linear piecewise growth mixture model with unknown knots: A comparison of two common approaches to inference. *Psychological Methods*, *20*, 259–275.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer.

Laird, N. M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, *65*, 581–590.

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*, 486–506. doi:10.1037/met0000075

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *The handbook of multivariate experimental psychology* (Vol. 2, pp. 561–614). New York, NY: Plenum Press.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 342–380). Lincolnwood, IL: Scientific Software International.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Measurement*, *14*, 126–149.

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 137–175). Washington, DC: American Psychological Association.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. New York, NY: Chapman & Hall.

Meng, X.-L., & van Dyk, D. (1997). The EM algorithm—An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society. Series B (Methodological)*, *59*, 511–567.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122.

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bifactor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*, 407–422.

Muthén, B., & Asparouhov, T. (2015). Growth mixture modeling with non-normal distributions. *Statistics in Medicine*, *34*, 1041–1058.

Muthén, B. O. (2010). *Bayesian analysis in Mplus: A brief introduction*. Retrieved from http://www.statmodel.com/download/introbayesversion%203.pdf

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

Paek, I., Li, Z., & Park, H. (2016). Specifying ability growth models using a multidimensional item response model for repeated measures categorical ordinal item response data. *Multivariate Behavioral Research*, *51*, 569–581.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.* Proceedings of the 3rd Internaitonal Workshop on Distributed Statistical Computing, March 20–22, Vienna, Austria.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika*, *73*, 167–182.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*, 51–67.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society (Series A)*, *158*, 73–90.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*. Retrieved from https://www.jstatsoft.org/article/view/v048i02

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*, 5–30.

Sheng, Y., & Wikle, C. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*, 413–430.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society (Series B)*, *64*, 583–639.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health outcomes. *Medical Care*, *44*, S39–S49.

Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, *59*, 225–255.

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.

Wang, C. (2014). Improving measurement precision of hierarchical latent traits using adaptive testing. *Journal of Equational and Behavioral Statistics*, *39*, 452–477.

Wang, C., Kohli, N., & Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 455–465.

Wang, C., & Nydick, S. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied Psychological Measurement*, *39*, 119–134.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*, 456–477.

Widaman, K. F., Ferrer, E., & Conger, R. (2010). Factor invariance within longitudinal structural equation models: measuring the same construct across time. *Child Development Perspectives*, *1*, 10–18.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Wilson, M., Zheng, X., & McGuire, L. W. (2012). Formulating latent growth using an explanatory item response model appraoch. *Journal of Applied Measurement*, *13*, 1–22.

Zhang, X., Tao, J., & Wang, C. (2019). Bayesian model selection methods for multilevel IRT models: A comparison of five DIC-based indices. *Journal of Educational Measurement*, *56*, 3–27.

Authors

CHUN WANG is an assistant professor of measurement and statistics in the College of Education at the University of Washington, 2012 Skagit Lane, Seattle, Washington, 98105; email: wang4066@uw.edu. Her research interests are multidimensional and multilevel item response theory, computerized adaptive testing, and cognitive diagnostic modeling.

STEVEN W. NYDICK is a data scientist developer at Korn Ferry, 33 South 6th St. #4900, Minneapolis, MN 55402; email: nydic001@umn.edu. His research interests include classification computerized adaptive testing, sequential testing in item response theory, efficiency/accuracy of parameter estimation algorithms, statistical programming, and statistical pedagogy.

Manuscript received February 17, 2017
First revision received November 22, 2017
Second revision received April 1, 2019
Accepted September 18, 2019