Journal of Educational and Behavioral Statistics
2014, Vol. 39, No. 4, pp. 235–256
DOI: 10.3102/1076998614531045

A Third-Order Item Response Theory Model for Modeling the Effects of Domains and Subdomains in Large-Scale Educational Assessment Surveys

Frank Rijmen
CTB/McGraw-Hill

Minjeong Jeon
Ohio State University

Matthias von Davier
Educational Testing Service

Sophia Rabe-Hesketh
University of California, Berkeley

Second-order item response theory models have been used for assessments consisting of several domains, such as content areas. We extend the second-order model to a third-order model for assessments that include subdomains nested in domains. Using a graphical model framework, it is shown how the model does not suffer from the curse of multidimensionality. We apply unidimensional, second-order, and third-order item response models to the 2007 Trends in International Mathematics and Science Study. Our findings suggest that deviations from unidimensionality are more pronounced at the content domain level than at the cognitive domain level and that deviations from unidimensionality at the content domain level become negligible after taking into account topic areas.

Keywords: *higher order factor analysis, multidimensional item response theory, graphical models*

Higher order factor analytic models are often used in research on the structure of cognitive abilities (e.g., Bickley, Keith, & Wolfe, 1995; Carroll, 1993; Taub, 2001). Thurstone (1944) introduced the idea of second-order factors, and Jöreskog (1970) gave a formal mathematical representation of the model. In a second-order model, the observed variables are assumed to satisfy a factor analytic model. The factors of this part of the model are called first-order factors. They are assumed to satisfy a factor analytic model themselves. That is, the correlations between the

first-order factors are modeled by a smaller set of second-order factors. Bentler (1976) extended the second-order model to a general higher order factor model by a recursive application of the same principle: The factors of order k are assumed to satisfy a factor analytic model of order $k + 1$.

Within the context of item response theory (IRT), de la Torre and Song (2009) proposed a second-order IRT model, which can be understood as the discrete response analog of a second-order factor model. They formulated the model within a Bayesian framework and estimated parameters using Markov chain Monte Carlo methods. In this article, we extend the model of de la Torre and Song to a third-order model. In addition, we show how both the second- and third-order structures allow for efficient full-information maximum likelihood estimation.

The field of educational measurement has seen relatively few applications of higher order IRT models. Therefore, we present several reasons for considering higher order models in the context of educational measurement models before proceeding to the technical part of this article. First, a higher order structure can be used to complement the measurement model with a model for the conceptual structure of the construct being measured. An empirical corroboration of the theoretical model for an assessment is part of the test validation process. In educational assessments, the assessment framework gives the theoretical basis for the assessment. An assessment framework provides a blueprint that determines the content to be assessed and guides item development. The experts who develop the assessment framework often adopt a hierarchical view of the assessment content. For example, the overall construct of mathematics in the Trends in International Mathematics and Science Study (TIMSS) is defined to encompass four content domains (e.g., algebra), each of which in turn consists of more narrowly defined topics (e.g., ''algebraic expressions'' and ''patterns''). Such a framework can be incorporated into the psychometric model by defining a corresponding higher order structure for its latent variables.

Second, higher order models can be useful in obtaining more reliable scores for domains or subdomains (de la Torre & Song, 2009). While there is a demand for more fine-grained information than overall scores, the reliability of subscores tends to be low if they are based on only a few items. Several methods have therefore been proposed to improve the reliability of subscores by taking into account the responses to the items of the other domains (de la Torre & Song, 2009; Haberman, 2008; Haberman & Sinharay, 2010; Wainer et al., 2001). De la Torre and Song (2009) pointed out that an advantage of using their second-order IRT model is that both overall scores and subscores can be obtained directly as expected a posterior (EAP) predictions for the and first- and second-order dimensions, respectively. This advantage becomes even more apparent when considering models with an order higher than 2. For example, in a third-order model, overall scores can be obtained as EAP predictions for the overall third-order dimension, domain scores as the predictions for the second-order dimensions, and

subdomain scores as predictions for the first-order dimensions. Expected total, domain, and subdomain sum scores can be obtained in a similar way.

A third reason for considering higher order models is that they nicely balance parsimony with model fit, allowing for multiple dimensions yet avoiding the rapid increase in the number of parameters for modeling the correlations between dimensions. In addition, in our experience, estimating a large unrestricted correlation matrix can become numerically unstable. This is especially the case when the correlations between first-order factors are high, as is typically found in educational assessments.

In the following, we present a set of higher order models that are aligned with the assessment framework of the TIMSS. First, we shortly describe TIMSS and the data sets that were used for the application.

TIMSS

TIMSS is an internationally comparative educational survey dedicated to improving teaching and learning in mathematics and science for students around the world (http://timssandpirls.bc.edu/TIMSS2007/about.html). The first study was carried out in 1995 and has been repeated every 4 years since then. In every cycle, fourth and eighth graders are assessed. In 2007, there were 59 participating countries. The data collected in the TIMSS 2007 are publicly available at http://timss.bc.edu/timss2007/idb_ug.html. In this article, we used the data of the 2007 eighth-grade mathematics assessment for the United States ($n = 7{,}377$).

The TIMSS international database lists 215 mathematics items for Grade 8 (http://timss.bc.edu/timss2007/idb_ug.html). All but one of those items were used in the item calibration in TIMSS. Each of the items pertains to one of three cognitive domains and also to one of four content domains. Cognitive domains specify the types of thinking processes that are assessed: *knowing*, *applying*, and *reasoning*. Content domains specify the subject matter that is assessed: *number*, *algebra*, *geometry*, and *data and chance*. Table 1 gives the contingency table of items in terms of cognitive and content domains. Within each content domain, items are further subdivided by topic areas. For example, the topic areas for algebra are *patterns, algebraic expressions,* and *formulas and functions*. Across the four content domains, there were 13 topic areas. The topic areas *whole numbers* and *integers* have been merged for reasons explained below. The classification of items into domains and topic areas was based on the judgments of subject matter experts.

Of the 214 items, 192 were dichotomous items, and 22 items had three score categories. Items were both of the multiple choice (116 items) and constructed response types (98 items). Forty-three items shared a common stimulus material with 1, 2, or 3 other items. This type of item clustering was not taken into account by the operational item calibration procedures, and neither will it be taken into account by the models presented in this article.

TABLE 1
Classification of TIMSS Mathematics Items in Terms of Content and Cognitive Domains

	Cognitive Domain			
Content	Knowing	Applying	Reasoning	Total
Algebra	32	15	17	64
Data and chance	14	18	8	40
Geometry	8	27	12	47
Number	27	28	8	63
Total	81	88	45	214

Note. TIMSS = Trends in International Mathematics and Science Study.

Like other large-scale assessments, participants in TIMSS only received a subset of the items (two blocks of mathematics items, with each block consisting of 11–18 items). Furthermore, participants are sampled according to a complex two-stage clustered sampling design. The sampling design calls for the use of sampling weights during model estimation, as recently discussed by Rutkowski, Gonzalez, Joncas, and von Davier (2010).

Higher Order IRT Models

Second-Order Model

The models presented all exhibit simple structure: Every item is allowed to load only on a single first-order dimension. For binary item responses, the second-order model of de la Torre and Song (2009) can be defined as follows. Let y_j denote the binary response on the jth item, $j = 1, \ldots, J$. Every item is embedded within an item cluster (content or cognitive domain, topic area) k, $k = 1, \ldots, K$. There are J_k items embedded within each item cluster; hence, $\sum_{k=1}^{K} J_k = J$. Furthermore, $\pi_j = P\left(y_j = 1 \middle| \theta_{k[j]}^{(1)}\right)$ is the probability of a correct response conditional on the latent variable $\theta_k^{(1)}$ associated with the item cluster k to which item j belongs (indicated by the subscript $k[j]$). The superscript "(1)" indicates that $\theta_k^{(1)}$ is a first-order dimension. The conditional response probability is related to a linear function of the latent variable through a link function $g(\cdot)$,

$$g(\pi_j) = \alpha_j^{(1)} \theta_{k[j]}^{(1)} - \beta_j, \tag{1}$$

where $g(\cdot)$ is typically the probit or logit link function. The parameter β_j is the intercept parameter and $\alpha_j^{(1)}$ is the slope or discrimination parameter of item j.

In this article, the logit link function is used for binary items, $g(\pi_j) = \log[\pi_j/(1-\pi_j)]$. For polytomous responses, the model can be extended in a straightforward way by choosing the cumulative logit link (as in graded response models) or the adjacent category logit link (as in partial credit models). For an item with response categories 1 to C_j, the cumulative logit link function is defined as $\log\left[\pi_j^{c+}\Big/\left(1-\pi_j^{c+}\right)\right]$, with $\pi_j^{c+} = \Pr(y_j > c|\theta)$ for $c = 1, \ldots, C_j - 1$. The adjacent category logit link function is defined as $\log\left[\pi_j^{c+1}\Big/\left(\pi_j^{c+1}+\pi_j^{c}\right)\right]$, with $\pi_j^{c} = \Pr(y_j = c|\theta)$ for $c = 1, \ldots, C_j - 1$. Both link functions reduce to the binary logit function when $C_j = 2$. In this article, the cumulative link function is used for polytomous items.

In a second-order model, the correlations between the latent variables $\theta_k^{(1)}$ are modeled by specifying a common second-order factor model for the first-order factors. Assuming a single second-order factor as in de la Torre and Song (2009), we obtain

$$\theta_k^{(1)} = \alpha_k^{(2)}\theta^{(2)} + \xi_k^{(1)}, \tag{2}$$

where $\theta^{(2)}$ represents the common second-order factor, $\alpha_k^{(2)}$ is the loading of the first-order factor k on the second-order factor, and $\xi_k^{(1)}$ is the random part of $\theta_k^{(1)}$ not accounted for by $\theta^{(2)}$. The second-order model formulation is completed by specifying univariate standard normal distributions for the latent variables $\theta^{(2)}, \xi_1^{(1)}, \ldots, \xi_k^{(1)}, \ldots, \xi_K^{(1)}$. Correlations between the first-order dimensions are induced by the second-order dimension:

$$\mathrm{Cov}\left(\boldsymbol{\theta}^{(1)}\right) = \boldsymbol{\alpha}^{(2)}\mathrm{Cov}\left(\theta^{(2)}\right)\boldsymbol{\alpha}^{(2)\prime} + \mathrm{Cov}\left(\xi^{(1)}\right) = \boldsymbol{\alpha}^{(2)}\boldsymbol{\alpha}^{(2)\prime} + \mathbf{I}_K, \tag{3}$$

where $\boldsymbol{\theta}^{(1)}$ is the vector of first-order dimensions, $\boldsymbol{\theta}^{(1)} = \left(\theta_1^{(1)}, \ldots, \theta_k^{(1)}, \ldots, \theta_K^{(1)}\right)'$; $\xi^{(1)}$ is the vector of first-order residual dimensions, $\xi^{(1)} = \left(\xi_1^{(1)}, \ldots, \xi_k^{(1)}, \ldots, \xi_K^{(1)}\right)'$; $\boldsymbol{\alpha}^{(2)}$ is the vector of loadings of first-order dimensions on the second-order dimension, $\boldsymbol{\alpha}^{(2)} = \left(\alpha_1^{(2)}, \ldots, \alpha_k^{(2)}, \ldots, \alpha_K^{(2)}\right)'$; and $\mathbf{I}_K$ is the K by K identity matrix.

The directed acyclic graph of a second-order model is presented in Figure 1. Boldface letters represent vectors, that is, $\mathbf{y}_k$ represents the response variables of all items belonging to item cluster k. A higher order model can always be expressed as a restricted hierarchical model by the use of a Schmid–Leiman transformation (Schmid & Leiman, 1957; Yung, Thissen, & McLeod, 1999). Indeed, combining Equations 1 and 2 gives

$$g(\pi_j) = \alpha_j^{(1)}\alpha_{k[j]}^{(2)}\theta^{(2)} + \alpha_j^{(1)}\xi_{k[j]}^{(1)} - \beta_j, \tag{4}$$

which is the expression of a bifactor model, with the restriction that the loadings on the general dimension $\theta^{(2)}$ are proportional to the loadings on the specific dimension $\xi_k^{(1)}$ within each item cluster k. Furthermore, the second-order model is formally equivalent to the testlet model of Bradlow, Wainer, and Wang (1999), which is usually expressed in a form similar to Equation 4 (Li, Bolt, & Fu, 2006; Rijmen, 2010). Figure 2 represents the directed acyclic graph for a bifactor model.

Three second-order models were specified for the 2007 TIMSS Grade 8 mathematics data. In the first model, the three cognitive domains were the first-order factors. In the second model, the four content domains were the first-order factors. Finally, the third model incorporated first-order factors for the topic areas. Because there are only three cognitive domains, the cognitive domain model is equivalent to a simple structure model with unconstrained correlations between the factors. The content domain model is slightly more constrained than the corresponding simple structure model: The six correlations between the four first-order factors are modeled with four second-order factor loadings. For the topic area model, however, the second-order model constrains the correlations substantially: 78 correlations are modeled with 13 loadings.

Third-Order Model

Because topic areas are clustered within content domains, one may expect that topic area dimensions from the same content domain will tend to be more highly correlated than dimensions from different content domains. If this is the case, the assumption that all correlations between topic area dimensions can be explained by a single second-order factor is violated. Specifying a third-order model is a natural way of relaxing this assumption. To our knowledge, the third-order item response model is a new model in psychometrics.

Figure 3 contains its directed acyclic graph. In the model, the relations between the conditional item response probabilities and a set of K first-order dimensions are specified as in Equation 1 (with $\theta_k^{(1)}$ representing topic area k). A second-order model is specified for the first-order dimensions. Instead of one common second-order dimension, there is a second-order dimension for each cluster of first-order dimensions. Specifically, for the third-order model that was defined for the 2007 TIMSS Grade 8 assessment, there is a second-order factor for each content domain,

$$\theta_k^{(1)} = \alpha_k^{(2)}\theta_{d[k]}^{(2)} + \xi_k^{(1)}, \tag{5}$$

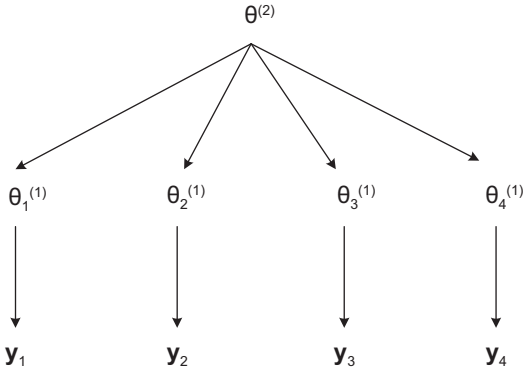

FIGURE 1. *Directed acyclic graph of a second-order model.*

where $\theta^{(2)}_{d[k]}$ represents the content dimension d in which topic area k is nested ($d = 1, \ldots, D$; $D = 4$ for 2007 TIMSS mathematics), and $\alpha^{(2)}_k$ is the loading of the first-order topic area factor k on the second-order content domain factor d. A common third-order factor accounts for the correlations among all second-order factors $\theta^{(2)}_d$,

$$\theta^{(2)}_d = \alpha^{(3)}_d \theta^{(3)} + \xi^{(2)}_d. \tag{6}$$

Analogous to the second-order model, independent univariate standard normal distributions are assumed for the latent variables $\theta^{(3)}, \xi^{(2)}_1, \ldots, \xi^{(2)}_D$, $\xi^{(1)}_1, \ldots, \xi^{(1)}_K$. The model-implied covariances between the second-order dimensions are

$$\mathrm{Cov}\left(\boldsymbol{\theta}^{(2)}\right) = \boldsymbol{\alpha}^{(3)}\boldsymbol{\alpha}^{(3)\prime} + \mathbf{I}_D, \tag{7}$$

where $\boldsymbol{\theta}^{(2)}$ is the vector of second-order dimensions, $\boldsymbol{\theta}^{(2)} = \left(\theta^{(2)}_1, \ldots, \theta^{(2)}_d, \ldots, \theta^{(2)}_D\right)'$; $\boldsymbol{\alpha}^{(3)}$ is the vector of loadings of second-order dimensions on the third-order dimension, $\boldsymbol{\alpha}^{(3)} = \left(\alpha^{(3)}_1, \ldots, \alpha^{(3)}_d, \ldots, \alpha^{(3)}_D\right)'$; and $\mathbf{I}_D$ is the D by D identity matrix. The covariance matrix of the first-order dimensions is given by

$$\mathrm{Cov}\left(\boldsymbol{\theta}^{(1)}\right) = \mathbf{A}^{(2)}\left(\boldsymbol{\alpha}^{(3)}\boldsymbol{\alpha}^{(3)\prime} + \mathbf{I}_D\right)\mathbf{A}^{(2)\prime} + \mathbf{I}_K, \tag{8}$$

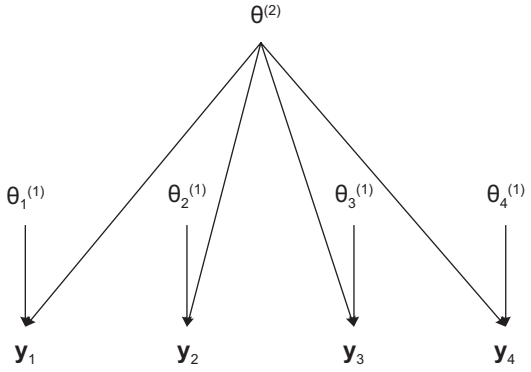

FIGURE 2. *Directed acyclic graph of a bifactor model.*

where $\mathbf{A}^{(2)}$ is the K by D loading matrix for the first-order dimensions, with elements $a_{kd} = \alpha_k^{(2)}$ if first-order cluster (e.g., topic) k is nested in the second-order cluster (e.g., content domain) d, and $a_{kd} = 0$ otherwise.

Combining Equations 5 and 6, we obtain

$$\theta_k^{(1)} = \alpha_k^{(2)}\alpha_{d[k]}^{(3)}\theta^{(3)} + \alpha_k^{(2)}\xi_{d[k]}^{(2)} + \xi_k^{(1)}, \tag{9}$$

and substituting Equation 9 into Equation 1 gives the reduced form

$$g(\pi_j) = \alpha_j^{(1)}\alpha_{k[j]}^{(2)}\alpha_{d[k]}^{(3)}\theta^{(3)} + \alpha_j^{(1)}\alpha_{k[j]}^{(2)}\xi_{d[k]}^{(2)} + \alpha_j^{(1)}\xi_{k[j]}^{(1)} - \beta_j, \tag{10}$$

where $d[k]$ is short for $d[k[j]]$. Hence, the third-order model is a restricted version of what could be called a trifactor model with the restriction that the loadings on the general dimension $\theta^{(3)}$ are proportional to the loadings on the more specific dimensions $\xi_d^{(2)}$ within each item cluster d at the domain level, and the loadings on the latter are in turn proportional to the loadings on the most specific dimensions $\xi_k^{(1)}$ within item cluster k at the subdomain level. Figure 4 contains the directed acyclic graph for a trifactor model.

Estimation

Maximum likelihood estimation of multidimensional IRT models involves integration over the space of all latent variables. In general, the integrals have no closed-form solution. Numerical integration over the joint space of all latent variables becomes computationally demanding, as the number of dimensions grows (Jeon, Rijmen, & Rabe-Hesketh, 2013; Rabe-Hesketh, Skrondal, & Pickles, 2005; von Davier & Sinharay, 2007). Specifically, when the integral is evaluated using Gaussian quadrature (e.g., Bock & Aitkin, 1981), the number of calculations involved

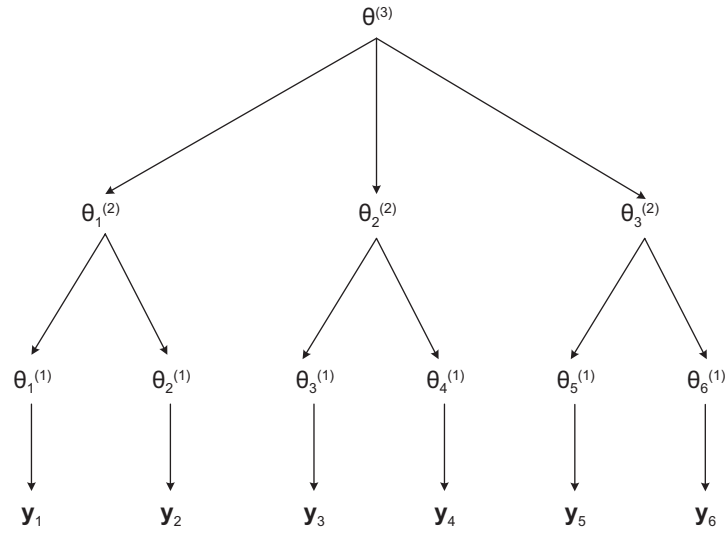

FIGURE 3. *Directed acyclic graph of a third-order model.*

increases exponentially with the number of latent variables. Even though the number of quadrature points per dimension can be reduced when using adaptive Gaussian quadrature (Pinheiro & Bates, 1995), the total number of points again increases exponentially with the number of dimensions. Furthermore, adaptive Gaussian quadrature involves the computation of the mode and the Hessian of the log posterior of the latent variables for each response pattern. The use of Monte Carlo techniques (e.g., stochastic expectation–maximization [EM], the Gibbs sampler in a Bayesian framework) has increased the number of dimensions that can be incorporated into a model, but for the high-dimensional models that are proposed, these techniques remain computation intensive (von Davier & Sinharay, 2007, 2010).

As an alternative, so-called limited information techniques have been developed in the field of structural equation modeling to deal with ordered categorical observed (indicator) variables (Jöreskog, 1994; Muthén, 1984). Unlike maximum likelihood estimation methods, the limited information techniques do not take into account the complete joint contingency table of all items, but only marginal tables up to the fourth order (Mislevy, 1985). In this way, parameter estimation can be carried out using weighted least squares estimation and is reasonably fast, even for high-dimensional models. By relying on marginal item frequency tables, limited information techniques are mostly suitable for complete data collection designs. Since all large-scale survey assessments employ an incomplete data collection design, however, limited information techniques cannot be used without further modifications in this context.

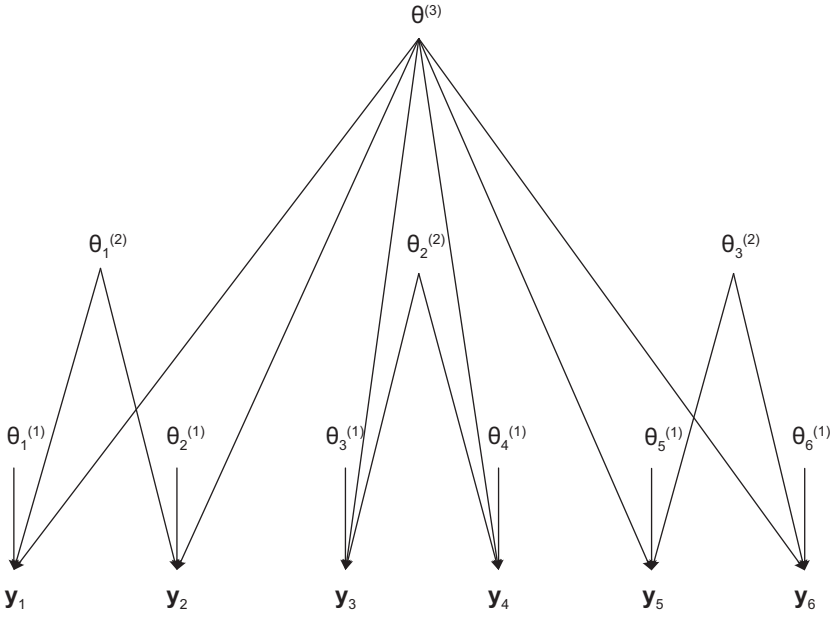

FIGURE 4. *Directed acyclic graph of a trifactor model.*

IRT models that do not incorporate item discrimination parameters, such as the Rasch (1960) model or the partial credit model (Masters, 1982), can be formulated as generalized linear mixed models (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). For those models, quasi-likelihood methods (Breslow & Clayton, 1993) have been developed as relatively fast alternatives to maximum likelihood estimation methods. Research showing that the early versions of these methods resulted in biased estimates has led to attempts to improve upon these methods (Goldstein & Rasbash, 1996; Raudenbush, Yang, Meng-Li, & Yosef, 2000). Notwithstanding the widespread use of limited information and quasi-likelihood estimation techniques and ongoing efforts for further improvements in these methods, one can assume that many researchers would prefer or at least consider using maximum likelihood estimation methods if the computational burden could be overcome.

Fortunately, maximum likelihood estimation is more feasible than it appears to be. A crucial realization is that the computational cost of maximum likelihood estimation is not necessarily driven by the number of dimensions. Depending on the conditional independence relations one is willing to assume, the actual computational cost can be far lower by exploiting these conditional relations during model estimation. In particular, the set of conditional independence relations implied by a model can be used to partition the joint space of all latent variables into smaller

subsets that are conditionally independent. The integral over the joint latent space can be evaluated numerically by defining grids over these smaller subsets of latent variables and carrying out a sequence of computations on these grids.

In general, the existence of more efficient integration methods follows from the fact that multiplication is distributive over addition. This rule can be applied repeatedly when computing the marginal probability of a response pattern, since the joint probability of the responses and latent variables is a product of factors, and the marginal probability is obtained by integrating the joint probability over the latent variables. Rearranging the order of integration and multiplication may result in a considerable reduction in the number of function evaluations. For example, for a bifactor or second-order model, the marginal probability of a response pattern can be obtained as follows:

$$\Pr(\mathbf{y}) = \int_{\theta^{(2)}} \int_{\xi_1^{(1)}} \cdots \int_{\xi_K^{(1)}} \phi\left(\theta^{(2)}\right) \left[\prod_k \Pr\left(\mathbf{y}_k \middle| \theta^{(2)}, \xi_k^{(1)}\right) \phi\left(\xi_k^{(1)}\right)\right] d\xi_K^{(1)} \ldots d\xi_1^{(1)} d\theta^{(2)}, \quad (11)$$

where $\phi(\cdot)$ denotes the standard normal distribution. First, we bring all factors that do not depend on $\xi_K^{(1)}$ outside of the integration over $\xi_K^{(1)}$:

$$\Pr(\mathbf{y}) = \int_{\theta^{(2)}} \int_{\xi_1^{(1)}} \cdots \int_{\xi_{K-1}^{(1)}} \phi\left(\theta^{(2)}\right) \left[\prod_k^{K-1} \Pr\left(\mathbf{y}_k \middle| \theta^{(2)}, \xi_k^{(1)}\right) \phi\left(\xi_k^{(1)}\right)\right] \int_{\xi_K^{(1)}} \Pr\left(\mathbf{y}_K \middle| \theta^{(2)}, \xi_K^{(1)}\right) \phi\left(\xi_K^{(1)}\right) d\xi_K^{(1)} \ldots, d\xi_1^{(1)} d\theta^{(2)}. \quad (12)$$

Then, we can approximate the rightmost integral numerically over $\xi_K^{(1)}$. The integrand contains only two latent variables, so that numerical integration can be carried out in a two-dimensional grid defined over $\xi_K^{(1)}$ and $\theta^{(2)}$. The result of integrating over $\xi_K^{(1)}$, $\Pr\left(\mathbf{y}_K \middle| \theta^{(2)}\right)$, depends only on one latent variable, $\theta^{(2)}$, and hence can be brought in front of the integrals over the other latent variables $\xi_K^{(1)}$, $k = 1, \ldots, K-1$:

$$\Pr(\mathbf{y}) = \int_{\theta^{(2)}} \Pr\left(\mathbf{y}_K \middle| \theta^{(2)}\right) \int_{\xi_1^{(1)}} \cdots \int_{\xi_{K-1}^{(1)}} \phi\left(\theta^{(2)}\right) \left[\prod_k^{K-1} \Pr\left(\mathbf{y}_k \middle| \theta^{(2)}, \xi_k^{(1)}\right) \phi\left(\xi_k^{(1)}\right)\right] d\xi_{K-1}^{(1)} \ldots, d\xi_1^{(1)} d\theta^{(2)}. \quad (13)$$

Repeating the same procedure for $k = K-1, \ldots, 1$,

$$\Pr(\mathbf{y}) = \int_{\theta^{(2)}} \phi\left(\theta^{(2)}\right) \left[\prod_k \Pr\left(\mathbf{y}_k \middle| \theta^{(2)}\right)\right] d\theta^{(2)}, \quad (14)$$

which can be numerically integrated over $\theta^{(2)}$. Hence, $\Pr(\mathbf{y})$ can be obtained through a sequence of computations in two-dimensional spaces.

Gibbons and Hedeker (1992) were the first to realize that maximum likelihood estimation for the bifactor model only requires function evaluations in two-dimensional spaces, but they did not explicitly refer to the property that multiplication is distributive over addition.

In principle, a similar approach could be followed for any other model, including the third-order and trifactor models discussed in the previous section. However, algebraic manipulations similar to the ones described in Equations 11 through 14 become cumbersome quite rapidly. In addition, it is not always straightforward to determine whether the dimensionality of the problem can be reduced, and neither is the optimal rearrangement of integrations and products always easy to determine. Thus, deriving efficient ways of computing the marginal likelihood for every sensible model and implementing them separately for each model does not seem to be a fruitful approach.

Fortunately, efficient ways of computing marginal probabilities can be obtained automatically and in a general way by adopting a graphical model framework. In a graphical model, random variables are represented by nodes, and conditional dependencies between random variables are represented by edges. A first advantage of graphical models is that a graph can be visualized easily using a diagram. For example, we have used diagrams of directed acyclic graphs in Figures 1 through 4 to represent the models presented in this article. In each of the figures, a directed edge from a parent node to a child node represents a direct conditional dependence of the random variable represented by the child node on the random variable represented by the parent node. For example, in Figure 1, the directed edge from $\theta_k^{(1)}$ to $\mathbf{y}_k$ represents that the item response variables of item cluster k depend on the first-order latent variable $\theta_k^{(1)}$. Visualization of model structures has been the primary use of graphs in the quantitative social and behavioral sciences.

The primary reason for the popularity of graphical models in machine learning and other research communities is that efficient methods of computing marginal probabilities can be obtained solely by manipulating the initial graphical representation of a statistical model. The core of the construction of efficient computational schemes relies on the transformation of the initial directed graphical representation of the model into an undirected triangulated graph and the subsequent construction of a junction tree. In a junction tree, the nodes correspond to subsets of variables. An important result is that those subsets of variables are conditionally independent of each other. As a consequence, the junction tree can be used to partition the high-dimensional space of all latent variables into subsets of lower dimensionality, and numerical integration over the joint (posterior) latent space can be carried out through a sequence of computations in these lower dimensional subspaces. The sequence in which computations have to be carried out is also provided by the junction tree. Again, all of this can be obtained by relying on algorithms defined on the

initial graphical representation of the statistical model. There is no need for tedious algebraic manipulations of the likelihood for each specific model. Readers with a further interest in graphical models are referred to Cowell, Dawid, Lauritzen, and Spiegelhalter (1999) for a more in-depth account. A very accessible introduction to graphical models can also be found in Bishop (2006, chap. 8).

The junction tree for the trifactor model represented in Figure 4 is given in Figure 5. It follows that maximum likelihood estimation for the trifactor model involves computations in three-dimensional subspaces formed by the general dimension $\theta^{(3)}$, one of the more specific dimension $\xi^{(2)}_{d[k]}$, and one of the most specific dimensions $\xi^{(1)}_{k}$. The third-order model was estimated as a trifactor model with proportionality restrictions as defined in Equation 10.

Maximum likelihood estimates can be obtained by using a modified EM algorithm (Lauritzen, 1995; Rijmen, Vansteelandt, & De Boeck, 2008). The E step of the modified algorithm is entirely embedded within a graphical model framework: Starting from the initial graphical representation of the model, a junction tree is constructed. The junction tree provides the sequence of lower dimensional subspaces in which computations are carried out during the E step. The modified EM algorithm is implemented as a toolbox in Matlab (Bayesian Networks with Logistic Regression Nodes [BNL]; Rijmen, 2006). As any other EM algorithm, the modified EM algorithm does not automatically provide standard errors as a by-product. Several procedures have been developed for obtaining standard errors when using the EM algorithm (McLachlan & Krishnan, 1997). In the BNL Matlab toolbox, there are two options to obtain standard errors depending on how the observed information matrix is approximated by the empirical information matrix (Meilijson, 1989) or by a numerical differentiation of the score function (which is routinely obtained in the M step of the EM algorithm). Currently, BNL does not take into account complex sampling designs in the computation of standard errors. Because large-scale educational surveys such as TIMSS employ a complex sampling design, standard errors are not reported in the application section. However, we did compute the observed information matrix based on numerical differentiation of the score function to verify model identification.

Application to the TIMSS 2007 Data

The following models were fitted: A unidimensional two-parameter logistic model (for comparison with the other models), three second-order models, respectively, with topic areas, content domains, and cognitive domains as primary dimensions, and a third-order model with topic areas as primary dimensions and content domains as secondary dimensions. For the unidimensional model, the integral over the latent variable was approximated numerically with Gaussian quadrature using 20 nodes. For the second- and third-order models, Gaussian quadrature with 10 nodes per dimension was used. We used a convergence criterion of 10^{-4} for

FIGURE 5. *Junction tree of a trifactor model.*

TABLE 2
Information Criteria for the Fitted Models

Model	Number of Parameters	Deviance	AIC[a]	BIC[b]
2PL	446	249,209	250,101	253,181
Second-order content	450	248,617	249,517	252,625
Second-order cognitive	449	249,012	249,910	253,011
Second-order topic	458	247,746	248,662	251,825
Third order	462	247,671	248,594	251,785

Note. 2PL = two-parameter logistic; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion. $n = 7{,}377$; $J = 212$.
[a]AIC = Deviance + $2 \times p$, with p the number of estimated parameters.
[b]BIC = Deviance + $\log(n) \times p$, with p the number of estimated parameters and n the sample size.

absolute differences in parameter estimates. We used maximum pseudolikelihood estimation, weighting the log likelihood contributions of students by the sampling weights that are provided with the TIMSS data for analyses that are carried out within countries (Foy & Olson, 2009).

A few of the items caused the estimation algorithm to exhibit convergence problems in one or more of the higher order models. Specifically, the discrimination parameter on one or more of the dimensions became very large for these items, resulting in numerical overflow. The problem items (M042301C and M042198C) both pertain to the patterns topic area (algebra content domain) and the reasoning cognitive domain. We decided to leave these items out of the model calibrations. Note that both items are items that shared a common stimulus material with other items (this is denoted by the capital A, B, or C at the end of the item identifier). Initially, the third-order model included separate topic area dimensions for whole numbers and integers. However, the information matrix evaluated at the parameter estimates was ill conditioned. Integers consisted of four items only, and the topic area is conceptually very similar to whole numbers. Therefore, we decided to combine both topic areas. Both the third-order model and the second-order topic area model were fitted again. The information matrices of the refitted models were no longer ill conditioned.

We used Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as informal criteria for model comparison. Both the AIC and the BIC were substantially lower for the second-order model than for the unidimensional model when either the topic areas or content domains were the first-order factors in the model (see Table 2).

Comparing the fit of each of the second-order models with the unidimensional model, the impact of incorporating topic areas as first-order dimensions was much larger than the impact of incorporating content domains as first-order dimensions, and the impact of incorporating cognitive domains as first-order factors was the smallest.

The third-order model had substantially lower AIC and BIC values than the second-order models with either cognitive or content domains, and slightly lower values on the AIC and the BIC than the second-order model with topic areas as first-order dimensions.

For all higher order models, the first- and second-order dimensions $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ were rescaled to have a variance of 1. The resulting standardized loadings are given in Table 3. The model-implied correlations between the content domains are presented in Table 4 for both the second-order model with content domains and the third-order model. In the second-order model with cognitive domains as first-order factors, the loadings of the cognitive domains on the second-order factor were very high (.98 and higher) for applying and knowing. The loading of reasoning was a bit lower. The loadings of the first-order factors tended to be a bit lower for the second-order model with content domains than in the model with cognitive domains, with model-implied correlations between content domains ranging from .84 to .89. However, the loadings of the content domains on the overall factor were consistently larger in the third-order model, with model-implied correlations between content domains of .92 and higher. Taken together, the results indicate that deviations from unidimensionality are more pronounced at the content domain level than at the cognitive domain level. However, the deviations from unidimensionality at the content domain level become negligible after taking into account the effects of topic areas, which are nested within content domains. This interpretation is consistent with the fact that the improvement in AIC and BIC values is relatively small when going from the second-order model with topic areas to the third-order model.

Parameter Recovery Study

A small simulation study was carried out to evaluate the finite sample properties of the marginal maximum likelihood estimators of the third-order model that was fitted in the previous section. Fifty data sets were generated using the maximum likelihood estimates as generating values. The number of items and persons was the same as the number of items and persons used in the application.

For each parameter, the bias and the root mean square error (RMSE) were estimated. Across the board, the individual item parameters were recovered well, with the median of the absolute value of the bias (across items) equal to 0.01 for both the discrimination ($\alpha_j^{(1)}$) and the intercept parameters (β_j) and the median of the RMSE equal to 0.04 for the intercept parameters equal to 0.02 for the discrimination parameters. The recovery of the higher order structure was of main

TABLE 3
Standardized Loadings of First (and Second)-Order Factors on Second- (and Third)-Order Factor

Model	First-Order Factor (Number of Items)	Second-Order Factor					Third-Order Factor
		Common	Number	Data and Chance	Algebra	Geometry	Common
Second-order content	Number (63)	.94					**.96**
Third-order (content on common)	Data and chance (40)	.90					**.98**
	Algebra (62)	.93					**.97**
	Geometry (47)	.95					**.95**
Second-order cognitive	Applying (88)	.98					
	Knowing (81)	1.00					
	Reasoning (43)	.89					
Second-order topic	Fractions and decimals (23)	.96	**1.00**				
Third-order (topic on content)	Ratio, proportion, and percentage (20)	.87	**.91**				
	Whole number and integer (20)	.91	**.93**				
	Chance (10)	.94		**.98**			
	Data org. and rep. (12)	.79		**.81**			
	Data interpretation (18)	.83		**.85**			
	Equations and function (25)	.94			**.98**		
	Algebraic expressions (21)	.92			**.94**		
	Patterns (16)	.68			**.69**		
	Location and mov. (9)	.95				**1.00**	
	Geometric measurement (12)	.93				**.99**	
	Geometric shapes (26)	.94				**.99**	

Note. The numbers in boldface are the loadings for the third-order model. Data org. and rep. = Data organization and representation; location and mov. = location and movement.

TABLE 4
Model-Implied Correlations Between the Cognitive Domains and Between the Content Domains

	Number	Data and Chance	Algebra	Geometry
Number		.95	.93	.92
Data and chance	.85		.95	.93
Algebra	.87	.84		.93
Geometry	.89	.86	.89	

Note. Elements below the diagonal are based on the second-order model with content domains; elements below the diagonal are based on the third-order model.

TABLE 5
Bias and Root Mean Squared Deviation (RMSE) for the Standardized Loadings of the First- and Second-Order Factors

Content Domain (Second Order)	Bias	RMSE
Number	−0.008	0.008
Data and chance	−0.011	0.011
Algebra	−0.013	0.013
Geometry	−0.008	0.009
Topic area (first order)	**Bias**	**RMSE**
Fractions and decimals	−0.002	0.002
Ratio, proportion, and percentage	−0.008	0.008
Whole number and integer	−0.006	0.007
Chance	0.001	0.001
Data org. and rep.	−0.013	0.014
Data interpretation	−0.010	0.011
Equations and function	0.007	0.008
Algebraic expressions	−0.007	0.008
Patterns	−0.017	0.018
Location and mov.	−0.004	0.004
Geometric measurement	0.002	0.002
Geometric shapes	−0.003	0.003

Note. $n = 50$. Data org. and rep. = Data organization and representation; location and mov. = location and movement.

interest in this study; estimated bias and RMSE are given in Table 5 for each standardized loading. The standardized loadings of the first-order factors on the second-order factors, as well as the standardized loadings of the second-order factors on the third-order factor, were recovered well. The absolute value of the estimated bias never exceeded 0.01, and the RMSE did not exceed 0.01 for all but

one standardized loading. Even though small, the bias for all four standardized loadings of the second-order factors was negative, indicating a small underestimation of the true standardized loadings for the second-order factors. For the first-order factors, both positive and negative bias was observed.

Subscores

De la Torre and Song (2009) showed that a second-order model can be used to obtain subscores that are more reliable than subscores that are based on subsets of items only. We verified that this was also the case for the third-order model by computing the reliability indices proposed by Haberman and Sinharay (2010) for multidimensional IRT models. In particular, we compared for each topic area the proportional reduction in mean square error (PRMSE) of the subscore based on a unidimensional IRT model that was calibrated to the items of the topic area only to the PRMSE of the subscore based on the multidimensional third-order model (Haberman & Sinharay, 2010, section 2.6). The PRMSEs are given in Table 6. The PRMSEs for the subscores based on a unidimensional IRT model are quite low, especially for the topic areas with a small number of items. In this regard, it is important to take into account that in TIMSS, every student receives only a small subset of the total item pool, so that the number of items per topic area administered to each individual student is very limited. The increase in PRMSE is quite large. Furthermore, the two topic areas with the lowest standardized loadings on the second-order dimensions had the lowest PRMSE based on the third-order model (but not based on the unidimensional IRT model). These findings are in line with de la Torre and Song (2009), who found the largest gain in reliability when using the second-order model for the case of a small number of items per domain and a high correlation between domains.

Concluding Remarks

We proposed the use of higher order multidimensional IRT models for the analysis of data from large-scale educational assessment surveys. Second-order multidimensional IRT models have been applied to educational assessment data by de la Torre and Song (2009). However, to our knowledge, we are the first to formulate and apply a third-order multidimensional IRT model. A possible explanation for why such a model has not been used before in the context of educational measurement is that maximum likelihood estimation of such a model seems computationally prohibitive at first sight. However, by taking advantage of the conditional independence relations of the model during estimation, maximum likelihood estimation is possible, even for a large number of items and a large number of persons. This was illustrated in the application using more than 200 items from the 2007 TIMSS mathematics assessment for Grade 8.

A second purpose of this article was to showcase the usefulness of graphical model theory as a framework for latent variable models. Graphical model theory

TABLE 6
Proportional Reduction in Mean Squared Error for Subscores Based on Separate Unidimensional IRT Models (PRMSE_U) and Based on the Third-Order Model (PRMSE_HO)

Topic Area (Number of Items)	PRMSE_U	PRMSE_HO
Fractions and decimals (23)	.47	.84
Ratio, proportion, and percentage (20)	.47	.77
Whole number and integer (20)	.44	.78
Chance (10)	.26	.83
Data org. and rep. (12)	.29	.63
Data interpretation (18)	.46	.72
Equations and function (25)	.50	.82
Algebraic expressions (21)	.43	.78
Patterns (16)	.46	.62
Location and mov. (9)	.16	.81
Geometric measurement (12)	.33	.80
Geometric shapes (26)	.38	.81

Note. IRT = item response theory; data org. and rep. = data organization and representation; location and mov. = location and movement.

can be used to construct efficient estimation methods in a general and automated way because it operates solely on the graphical representation of the model. No analytical derivations that are specific to a single model are required. In this article, graphical model theory was used for higher order multidimensional IRT models. For its application to psychometric models with discrete latent variables, such as higher order latent class models and latent Markov models, see Rijmen (2011) and Rijmen, Vansteelandt, and De Boeck (2008), respectively.

The price one has to pay for efficient maximum likelihood estimation methods is that one has to impose certain conditional independence assumptions on the latent structure of the model. However, by making these assumptions, we can take into account many more sources of individual differences. At a practical level, the real choice is between a model with a small number of dimensions with no further structure imposed and a rich model with a potentially very large number of dimensions that does incorporate some assumptions of conditional independence between dimensions. We think the latter is to be preferred, especially given the fact that these conditional independence assumptions often can be derived from the underlying theoretical assessment framework. In addition, it is common practice to build some modularity into complex statistical models through the use of conditional independence relations. For example, multilevel models incorporate the assumption that Level 1 units are independent conditional on the Level 2 units. In that sense, second- and third-order models are related to multilevel IRT models (e.g., Fox & Glas, 2001), the difference being that the former take into account a hierarchical structure at the item level, and the latter a hierarchical structure at the person level.

Currently, a multidimensional IRT model with separate dimensions for content domains is in operational use for TIMSS. When only considering the second-order model with content domains as first-order factors, the results do suggest that there is a unique contribution of content domains on top of a strong common dimension for mathematics. However, because topic areas are clustered within content domains, the second-order model cannot be used to determine whether these effects are due to the content domains per se or should be attributed to the effects of topic areas. The third-order model does allow for such a distinction, and the results indicate that the unique contribution of the content domain level vanishes after taking into account topic areas.

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: The research reported in this article was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R305D110027 to Educational Testing Service. The opinions expressed are those of the authors and do not represent the views of the Institute or the Department of Education.

References

Bentler, P. M. (1976). Multistructure statistical model applied to factor analysis. *Multivariate Behavioral Research*, *11*, 3–25.

Bickley, P. G., Keith, K. Z., & Wolfe, L. M (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence*, *20*, 309–328.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer-Verlag.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 443–459.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

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

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, England: Cambridge University Press.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. New York, NY: Springer.

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.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288.

Foy, P., & Olson, A. M. (2009). *TIMSS 2007 international database and user guide*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

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

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

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, *38*, 32–60.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, *57*, 239–251.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381–389.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, *19*, 191–201.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*, 3–21.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York, NY: Wiley.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, B*, *51*, 127–138.

Mislevy, R. J. (1985). *Recent developments in the factor analysis of categorical variables* (ETS Research Rep. No. RR-85-24). Princeton, NJ: ETS.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Pinheiro, P. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*, 12–35.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*, 301–323.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9*, 141–157.

Rijmen, F. (2006). *BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes* (Technical Report). Amsterdam, the Netherlands: VU University Medical Center.

Rijmen, F. (2010). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second order multidimensional IRT model. *Journal of Educational Measurement*, *47*, 361–372.

Rijmen, F. (2011). The latent class model as a measurement model for situational judgment tests. *Psychologica Belgica*, *51*, 197–212.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205.

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

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*, 142–151.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.

Taub, G. E. (2001). A confirmatory analysis of the Wechsler adult intelligence scale-third edition: Is the verbal/performance discrepancy justified? *Practical Assessment, Research & Evaluation*, *7*. Retrieved from http://PAREonline.net/getvn.asp?v=7&n=22

Thurstone, L. L. (1944). Second-order factors. *Psychometrika*, *9*, 71–100.

von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics*, *32*, 233–251.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation for latent regression item response models. *Journal of Educational and Behavioral Statistics*, *35*, 174–193.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., & Nelson, L. (2001). Augmented scores—"Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Hillsdale, MI: Erlbaum.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher order factor model and the hierarchical factor model. *Psychometrika*, *64*, 113–128.

Authors

FRANK RIJMEN is a principal research scientist and research manager at CTB/McGraw-Hill, Monterey, CA; e-mail: frank.rijmen@ctb.com. His primary research interest include latent variable modeling, graphical models, and item response theory.

MINJEONG JEON is an assistant professor in quantitative psychology at the Ohio State University, OH; e-mail: jeon.117@osu.edu. Her primary research interests include item response, multilevel, and latent variable modeling.

MATTHIAS VON DAVIER is a codirector at the Center for Global Assessment, Educational Testing Service, Princeton, NJ; e-mail: mvondavier@ets.org. His primary research interests are latent variable modeling including item response models, latent structure models, and mixture distribution models.

SOPHIA RABE-HESKETH is a professor at the Graduate School of Education and Graduate Group in Biostatistics, University of California, Berkeley, CA; e-mail: sophiarh@-berkeley.edu. Her primary research interests include multilevel and latent variable modeling and missing data.

Manuscript received August 7, 2012
Revision received November 25, 2013; February 25, 2014
Accepted March 17, 2014