Bayesian Model Selection Methods for Multilevel IRT Models: A Comparison of Five DIC-Based Indices

Xue Zhang and Jian Tao
Northeast Normal University
Chun Wang
University of Washington
Ning-Zhong Shi
Northeast Normal University

The project is supported by Key Laboratory of Applied Statistics of MOE, National Natural Science and Social Science Foundations of China (Grants 11571069) and Institute of Education Sciences grant R305D170042 (originally R305D160010).

Citation: Zhang, X., Tao, J., & Wang, C, & Shi, N. (2019). Bayesian model selection methods for multilevel IRT models: A comparison of five DIC-based indices. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12197

JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Spring 2019, Vol. 56, No. 1, pp. 3–27

Bayesian Model Selection Methods for Multilevel IRT Models: A Comparison of Five DIC-Based Indices

Xue Zhang and Jian Tao
Northeast Normal University
Chun Wang
University of Washington
Ning-Zhong Shi
Northeast Normal University

Model selection is important in any statistical analysis, and the primary goal is to find the preferred (or most parsimonious) model, based on certain criteria, from a set of candidate models given data. Several recent publications have employed the deviance information criterion (DIC) to do model selection among different forms of multilevel item response theory models (MLIRT). The majority of the practitioners use WinBUGS for implementing MCMC algorithms for MLIRT models, and the default version of DIC provided by WinBUGS focused on the measurement-level parameters only. The results herein show that this version of DIC is inappropriate. This study introduces five variants of DIC as a model selection index for MLIRT models with dichotomous outcomes. Considering a multilevel IRT model with three levels, five forms of DIC are formed: first-level conditional DIC computed from the measurement model only, which is the index given by many software packages such as WinBUGS; second-level marginalized DIC and second-level joint DIC computed from the second-level model; and top-level marginalized DIC and top-level joint DIC computed from the entire model. We evaluate the performance of the five model selection indices via simulation studies. The manipulated factors include the number of groups, the number of second-level covariates, the number of top-level covariates, and the types of measurement models (one-parameter vs. two-parameter). Considering the computational viability and interpretability, the second-level joint DIC is recommended for MLIRT models under our simulated conditions.

Model selection is important in any model-based inference. Taking item response theory (IRT) model as an example, the selection of a misspecified model leads to not only theoretically different interpretations of the data but also inappropriate conclusions with respect to other IRT applications such as biased parameter estimation, differential item functioning (DIF), or inappropriate person-fit assessment (DeMars, 2010).

The ease of fitting hierarchical models using Markov chain Monte Carlo (MCMC) algorithm has facilitated the development of model selection criterion within Bayesian framework. Congdon (2003) provided versions of Akaike's (1974) information criterion (AIC) and Schwarz's (1978) Bayesian information criterion (BIC)

Areas of Specialization: Statistics, Quantitative/Psychometric Methods.
Areas of Specialization: Statistics.

© 2019 by the National Council on Measurement in Education

when the fully Bayesian estimation methods, such as the MCMC algorithm, are used. Both AIC and BIC are based on the likelihood with a penalty, and their difference lies on the penalty term, which depends on the effective number of parameters in the model. This effective number is a measure of model complexity, which is often difficult to calculate for hierarchical models. This is because, although the number of parameters follows directly from the likelihood, the prior distribution imposes additional restrictions on the parameter space and it reduces the effective dimension (Entink, Fox, & van der Linden, 2009; Fox, 2010).

Within the Bayesian framework, a common approach for comparing two models is to compute the Bayes factor (BF; Berger & Delampady, 1987; Gelfand, 1996; Jeffreys, 1961; Kass & Raftery, 1995), which is defined as the ratio of the posterior probabilities of two models given data. Supposing that the prior densities of both models consist of a point mass at their respective MLEs (maximum likelihood estimates), and replacing the posterior probabilities by the likelihood of the two model parameters evaluated at their respective MLEs, the Bayesian factor becomes the classical likelihood ratio (Ando, 2010). Kass and Wasserman (1995) showed that under certain conditions the BIC was an approximation of the BF. An advantage of the BF is its clear interpretation of the change in the odds in favor of the model when moving from the prior to the posterior distribution (Lavine & Schervish, 1999). Unfortunately, BF is quite difficult both to compute and to interpret for high-dimensional hierarchical models and for models having improper prior distributions.

As a remedy, Geisser and Eddy (1979) discussed cross-validation in Bayes regression model comparison and proposed a so-called pseudo-Bayes factor (PsBF). The PsBF uses a "leave-one-out" method to calculate the cross-validation predictive densities (Gelfland & Dey, 1994) so that it can avoid intractable computation and dependence on the prior. The pseudo-marginal likelihood used here may be interpreted as a predictive measure for a future replication of the given data (Ando & Tsay, 2010). The PsBF is provided by the ratio of two such quantities from the two competing models. This predictive density yields the conditional predictive ordinates (CPO) index such that PsBF can be expressed as the ratio of two CPO indices. One drawback of PsBF, as noted by Eklund and Karlsson (2007), is that the division of the data into subsets may affect the results. Yet there exist no clear guidelines for the division, and the approach is difficult to apply when the data are dependent, as in, for instance, time series data. When the number of observations is large, the approach consumes a substantial amount of computational time (Ando & Tsay, 2010).

In addition to BF, Spiegelhalter, Best, Carlin, and Van Der Linde (2002) developed the deviance information criterion (DIC) as a measure of global model fit, which is computed based on Bayesian posterior estimates of model parameters. DIC is usually viewed as the Bayesian counterpart of AIC, which is approximately equivalent to AIC for models with negligible prior information, and it is easily obtained as a byproduct of the MCMC sampling algorithm. Further, it also makes weaker assumptions and automatically penalizes model complexity (Bolker et al., 2009). Despite these advantages, there still exist many weaknesses of DIC, as discussed by Spiegelhalter et al. (2002) and comments therein and by a series of articles by Celeux, Forbes, Robert, and Titterington (2006), Carlin (2006), Meng and Vaida (2006), and Plummer (2006). Because DIC is widely used for hierarchical/multilevel models, the

main criticism is that DIC is dependent on the level of parameter specification upon which the model likelihood is conditioned (i.e., "parameter of focus") and hence lacks invariance to reparameterization. Millar (2009) presented three variants of DIC for a three-level hierarchical model: (1) the conditional DIC, which used likelihood conditioned on parameters at the lowest level of the hierarchy, (2) the second-level marginalized DIC, which used partially marginal likelihood conditioned on parameters at the first and second levels, and (3) the top-level marginalized DIC, which used marginal likelihood conditioned on parameters at all levels. Hamaker, van Hattum, Kuiper, and Hoijtink (2011) introduced the conditional DIC and marginal DIC; the marginal DIC is similar to Millar's second-level marginalized DIC.

This article focuses on the model selection for the family of multilevel IRT models (e.g., Goldstein, 2003; Raudenbush & Bryk, 2002), which are commonly used to model nested structures in behavioral and social sciences with categorical outcomes. Many researchers have used different evaluation criteria to evaluate the global fit of multilevel models. For example, Entink et al. (2009) have used the conditional DIC provided in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003) to assess the global fit of a multilevel item response theory model (MLIRT), and other similar applications include Hamaker et al. (2011), Hung and Wang (2012), and Choi and Wilson (2016). Fox (2010), in contrast, used the top-level marginalized DIC to assess the fit of a MLIRT model with an application to the PISA (Program for International Student Assessment) data. Geering, Glas, and van der Linden (2011) and Wang, Chang, and Douglas (2013) used DIC calculated from the joint likelihood (without any marginalization) for a linear item cloning model and semi-parametric hierarchical linear transformation model, respectively. This type of DIC, however, does not fall into any of the three variants of DIC described by Millar (2009).

Furthermore, Hung and Wang (2012) have used BF to compare the generalized multilevel facets model for longitudinal data, and BF was also used by Entink et al. (2009). Choi and Wilson (2016) used the posterior predictive model check (PPMC; Gelman, Carlin, Stern, & Rubin, 1996; Guttman, 1967) method to investigate the effect of incorrect modeling school membership in the analysis of multilevel and longitudinal item response data.

Other information-based indices, such as the average of AIC and BIC, which were computed from the post burn-in iterations of MCMC algorithms, were used by Cho and Cohen (2010); AIC and the quasi-information criterion (QIC) were used by Barnett, Koper, Dobson, Schmiegelow, and Manseau (2010) and compared with DIC, and they recommended DIC; conditional AIC computed from the measurement model, marginal AIC computed from the measurement model and second level model, and BIC were used by Hamaker et al. (2011).

As noted by Hamaker et al. (2011), more often than not the selection of a model fit evaluation criterion is based on the capacity of the software. For example, the DIC provided in WinBUGS by default is a conditional DIC computed from the level-1 model, whereas the AIC and BIC from MLwiN (Rasbash et al., 2000), SPSS, R, and Mplus (Muthén & Muthén, 2010) are based on the marginal likelihood computed from the first and the second-level models.

Studies related to evaluating the performance of different variants of DIC are scarce. To the authors' best knowledge, there is only one study (Millar, 2009), that

used three variants of marginalized DIC to compare the hierarchical Bayesian models on over-dispersed count data, and the second-level marginalized DIC was recommended. However, this recommendation may not hold for binary item response data. Moreover, there were no systematic investigations comparing the performance of marginalized DIC and joint DIC. Without integrating out random effects at different levels of hierarchy, the joint DIC, in theory, should be computationally much easier than the marginalized DIC. It is yet to be evaluated in the present study to check if joint DIC can still select the best-fitting model under various conditions.

The current study focuses on multilevel IRT (MLIRT) models and is designed to investigate the performance of five variants of DIC: the first-level conditional DIC (DIC_C), the second-level marginalized DIC (DIC_S), the second-level joint DIC ($\mathrm{DIC}j_S$), the top-level marginalized DIC (DIC_T), and the top-level joint DIC ($\mathrm{DIC}j_T$). The outline of the rest of the article is as follows. In Section 2, we briefly review the multilevel IRT models. In Section 3, we present five variants of DIC as model selection criteria for MLIRT models. Simulation studies are provided in Section 4 to illustrate the performance of these five variants of DIC. A real-data analysis is provided in Section 5. We end with some concluding remarks in Section 6.

Model Description

In this section, we give a brief overview of the MLIRT model that is considered in this article. Interested readers can refer to Fox (2010) for a full description of the family of MLIRT models. The MLIRT model consists of two components.

Measurement Model: Level 1

In this article, the probit link is considered as the linking function, so that the posterior distributions of item parameters and latent trait (i.e., ability) have closed forms, which facilitate computing marginal likelihoods. And the two-parameter normal ogive (2PNO) model and one-parameter normal ogive (1PNO) model (Baker & Kim, 2004; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Lord, 1980) are considered as the measurement models, because Kang and Cohen (2007) reported that DIC did not work well with data from the three-parameter models. Let y_{ijk} denote the binary scored response of examinee i ($i = 1, \ldots, n_j$) in group j ($j = 1, \ldots, J$) on item k ($k = 1, \ldots, K$); the probability of a correct response is given by

$$P(y_{ijk} = 1 \,|\theta_{ij}, a_k, b_k) = \Phi(a_k\theta_{ij} - b_k), \tag{1}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function, and a_k and b_k are the discrimination and difficulty parameters of item k, set $a_k = 1$ when 1PNO model is used. θ_{ij} denotes the ability of examinee i in group j. Hereafter, the parameters of item k will also be succinctly denoted by a 2-by-1 vector, $\boldsymbol{\xi}_k$, that is, $\boldsymbol{\xi}_k = (a_k, b_k)^T$.

Structural Multilevel Model: Level 2 and Level 3

The structural multilevel model explains the relations between the latent variables and other observed variables:

Level 2

$$\theta_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \cdots + \beta_{qj}x_{qij} + \cdots + \beta_{Qj}x_{Qij} + e_{ij}, \tag{2}$$

Level 3

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}\omega_{1j} + \cdots + \gamma_{0S}\omega_{Sj} + u_{0j},\\
\beta_{1j} &= \gamma_{10} + \gamma_{11}\omega_{1j} + \cdots + \gamma_{1S}\omega_{Sj} + u_{1j},\\
&\vdots\\
\beta_{Qj} &= \gamma_{Q0} + \gamma_{Q1}\omega_{1j} + \cdots + \gamma_{QS}\omega_{Sj} + u_{Qj}.
\end{aligned} \tag{3}$$

In level 2, x_{qij} denotes the qth ($q = 0, \ldots, Q$) individual-specific covariate of examinee i in group j, such as socioeconomic status or gender. β_{qj} is the corresponding regression coefficient, and e_{ij} denotes the random effect at an individual level and is assumed to follow a normal distribution with a constant variance, that is, $e_{ij} \sim N(0, \sigma^2)$. In level 3, ω_{sj} denotes the sth ($s = 0, \ldots, S$) school-specific covariate of group j, such as teacher satisfaction or school climate. γ_{qs} is the corresponding regression coefficient, u_{qi} denotes the random effect at the school level and $\mathbf{u}_j \sim N(0, \mathbf{T})$, where $\mathbf{T} = \begin{bmatrix} \tau_{00}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \tau_{QQ}^2 \end{bmatrix}$ is a Q-by-Q covariance matrix.

Some restrictions are imposed to remove the scale indeterminacy inherent in normal ogive models. Two sets of constraints are usually adopted in the literature. One is to fix the scale of the ability to a standard normal distribution. As a result, the structural multilevel IRT model in Equations 2 and 3 is identified owing to the fixed scale of the abilities. Another way is to put a restriction on the item parameters, which can be accomplished by imposing the restriction $a_1 = 1$ and $b_1 = 0$. In this article, we take the second approach.

The Gibbs sampler (Albert, 1992; Fox & Glas, 2001) is used for MLIRT model estimation. The details are provided in Appendix 1.

Model Selection Methods

In this section, we introduce five variants of the DIC of Spiegelhalter et al. (2002) that will be used for hierarchical models. The generic form of DIC (Spiegelhalter et al., 2002) is expressed as

$$\mathbf{DIC} = 2\,\overline{D(\Theta)} - D(\bar{\Theta}), \tag{4}$$

where $D(\Theta) = -2\log\{p(\mathbf{y}|\Theta)\} + 2\log\{p(\mathbf{y})\}$ denotes Bayesian deviance, and the overline denotes posterior expectation. In Equation 4, $\mathbf{y}$ is the data matrix, and the deviance is conditional on parameter vector $\boldsymbol{\Theta}$, which is termed the "parameter of interest" or "parameter in focus" by Spiegelhalter et al. (2002). Here, it is sufficient to assume that the standardizing factor $p(\mathbf{y})$ equals one such that $D(\Theta) = -2\log\{p(\mathbf{y}|\Theta)\}$ (Fox, 2010). The number of effective model parameters p_D equals $\overline{D(\Theta)} - D(\hat{\Theta})$. Finally, the best-fitting model is associated with the smallest DIC value.

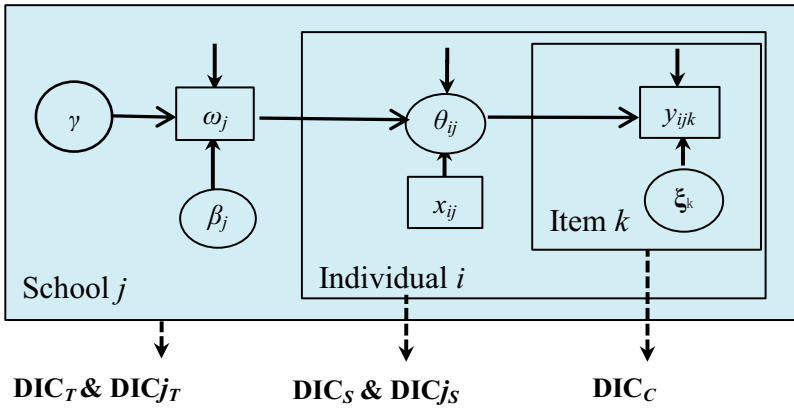

Figure 1. An illustration of five variants of DIC from a MLIRT model. (Color figure can be viewed at wileyonlinelibrary.com)

Figure 1 shows the parameters of focus in these five variants of DIC from a MLIRT model. Three boxes are plotted, implying that items are nested within individuals, whereas individuals are nested within schools. Different variants of DIC are marked graphically in this figure. For example, DIC_C only focuses on the lowest measurement model level, whereas DIC_S and $\text{DIC}j_S$ focus on the second level, and the last two focus on the top level. Observed variables are in squares, latent variable and parameters are in circles.

As the Gibbs sampler is used to fit the MLIRT model, after augmenting the discrete data **y** with the continuous data **z,** the DIC in Equation 4 can be expressed as the integrated augmented DIC; that is,

$$\begin{aligned}\text{DIC} &= \int [\text{DIC}\,|\mathbf{z}, \Theta\,] \cdot p\,(\mathbf{z}, \Theta\,|\mathbf{y}\,)\, d\mathbf{z} d\Theta \\ &= \int \left[2\overline{D(\mathbf{z}, \Theta)} - D\left(\mathbf{z}, \hat{\Theta}\right)\right] \cdot p\,(\mathbf{z}, \Theta\,|\mathbf{y}\,)\, d\mathbf{z} d\Theta \\ &= \int \left\{-4E_{\Theta|Z}\,\{\log[\,p\,(\mathbf{z}\,|\Theta\,)]\} + 2\log\left[p\left(\mathbf{z}\,\middle|\hat{\Theta}\right)\right]\right\} \cdot p\,(\mathbf{z}, \Theta\,|\mathbf{y}\,)\, d\mathbf{z} d\Theta, \quad (5)\end{aligned}$$

where $p(\mathbf{z}|\boldsymbol{\Theta})$ is the augmented likelihood function, $\boldsymbol{\Theta}$ is the matrix of the parameters, and $\hat{\Theta}$ is the point estimated value based on MCMC samples. z_{ijk} can be sampled from the full conditional posterior density:

$$z_{ijk}\big|\,\theta, \xi, \mathbf{y} \sim \begin{cases} N(a_k\theta_{ij} - b_k, 1) \text{ truncated at the left by 0 if } y_{ijk} = 1 \\ N(a_k\theta_{ij} - b_k, 1) \text{ truncated at the right by 0 if } y_{ijk} = 0 \end{cases}. \quad (6)$$

The detailed calculations of the augmented likelihood for five versions of DIC are presented in Appendix 2. Hereafter, we focus the definition calculation of $D(\mathbf{z}, \boldsymbol{\Theta})$, instead of $D(\boldsymbol{\Theta})$.

In the present context, it is computationally most convenient to take $\boldsymbol{\Theta}$ as the parameters in the lowest (i.e., measurement) level of the hierarchy, as specified by

Equation 1. This will be referred to as the conditional model and the corresponding deviance is

$$D(\mathbf{z}, \theta, \xi) = -2\log\{p(\mathbf{z}\,|\theta, \xi)\}. \tag{7}$$

The DIC calculated using the deviance defined in Equation 7 is the first-level conditional DIC (DIC_C; Millar, 2009). The first-level conditional DIC is defined similarly as the DIC used by Entink et al. (2009) and Fox (2010). Millar (2009) mentioned that the first-level conditional DIC ignored the immediate information provided by the higher-level structures, although the estimates of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ may still carry the higher-level information indirectly, therefore it was not sensitive to differentiate models that differ in higher levels. Even so, given its computational ease, it is still widely used and it is actually the default option in WinBUGS. Also note that the first-level conditional DIC is built on the "complete-data" likelihood from level 1 model, assuming random effect at level 1 (e.g., $\boldsymbol{\theta}$) is known by plugging in their estimated values. In so doing, it treats the estimated random effects as fixed values when computing DIC.

Furthermore, the parameters at the second level can sometimes be of interest when one intends to evaluate whether certain individual level (i.e., level 2) covariates have significant effects. Both marginal likelihood and joint likelihood are considered here. The second-level marginalized deviance $D(\mathbf{z}, \xi, \beta, \sigma^2)$ is defined as

$$D(\mathbf{z}, \xi, \sigma^2, \beta) = -2\log\{p(\mathbf{z}\,|\xi, \sigma^2\,, \beta)\}. \tag{8}$$

Fox's method (Fox, 2010) for calculating marginalized Bayesian deviance and Chib's method (Chib, 1995) for calculating marginal densities are applied here. The idea is to obtain a closed-form expression of the marginal likelihood using Bayes's formula. Hence, the corresponding second-level marginalized distribution has the following density:

$$\begin{aligned} p(\mathbf{z}\,|\xi, \sigma^2, \beta) &= \int p(\mathbf{z}\,|\xi, \theta)p(\theta\,|\sigma^2, \beta)d\theta \\ &= \frac{p(\mathbf{z}, \theta\,|\xi, \sigma^2, \beta)}{p(\theta\,|\mathbf{z}, \xi, \sigma^2, \beta)} = \frac{p(\mathbf{z}\,|\xi, \theta)p(\theta\,|\sigma^2, \beta)}{p(\theta\,|\mathbf{z}, \xi, \sigma^2, \beta)}. \end{aligned} \tag{9}$$

The three parts on the right side of Equation 9 all have closed forms and the full computation details are provided in Appendix 2. The DIC calculated using the deviance in Equation 8 is the second-level marginalized DIC (DIC_S; Millar, 2009). As compared to the complete-data likelihood used in the first-level conditional DIC, this form of DIC uses the second-level marginalized likelihood (see Equation 9), which can be viewed as an observed likelihood from model level 2. The "missing" data of $\boldsymbol{\theta}$ is marginalized, and hence the second-level marginalized DIC belongs to the family of "observed" DICs (Celeux et al., 2006).

The second-level joint deviance D_j ($\mathbf{z}$, $\boldsymbol{\xi}$, σ^2, $\boldsymbol{\beta}$) is defined as

$$D_j\left(\mathbf{z}, \xi, \sigma^2, \beta\right) = -2\log\left\{p\left(\mathbf{z}, \theta\,|\xi, \sigma^2, \beta\right)\right\}, \tag{10}$$

and then the corresponding second-level joint distribution is given by

$$p\left(\mathbf{z}, \theta \left| \xi, \sigma^2, \beta \right.\right) = p\left(\mathbf{z} \left| \xi, \theta \right.\right) p\left(\theta \left| \sigma^2, \beta \right.\right). \tag{11}$$

The DIC calculated using the deviance in Equation 10 will be called the second-level joint DIC (DICj_S), which can be treated as one kind of complete DIC (Celeux et al., 2006) because it is again based on complete-data likelihood. The unobserved, missing random effects are treated as known. Actually it is a semi-complete DIC because there is no immediate information from the highest level, that is, level 3.

It is also possible to calculate a top-level marginalized DIC or top-level joint DIC when the interest is focused on the significant school level (i.e., level 3) effects. The top-level marginalized deviance is given by

$$D\left(\mathbf{z}, \xi, \sigma^2, \gamma, \mathbf{T}\right) = -2 \log\left\{p\left(\mathbf{z} \left| \xi, \sigma^2, \gamma, \mathbf{T} \right.\right)\right\}. \tag{12}$$

Based on the Bayes' formula, the corresponding density can be expressed as

$$\begin{aligned} p(\mathbf{z}|\xi, \sigma^2, \gamma, \mathbf{T}) &= \int p(\mathbf{z}|\xi, \sigma^2, \beta) p(\beta|\gamma, \mathbf{T}) d\beta \\ &= \iint p(\mathbf{z}|\xi, \theta) p(\theta|\sigma^2, \beta) p(\beta|\gamma, \mathbf{T}) d\theta d\beta \\ &= p(\mathbf{z}|\xi, \sigma^2, \beta, \gamma, \mathbf{T}) \frac{p(\beta|\gamma, \mathbf{T})}{p(\beta|\mathbf{z}, \xi, \sigma^2, \gamma, \mathbf{T})}. \end{aligned} \tag{13}$$

The three parts on the right side of Equation 13 all have closed forms, and the computation details will also be provided in Appendix 2. The DIC calculated using the deviance in Equation 12 is the top-level marginalized DIC (DIC$_T$; Millar, 2009). Similarly, the top-level marginalized DIC is another version of observed DIC (Celeux et al., 2006), which can be calculated using an observed likelihood from model level 2 and level 3. The "missing" data of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are integrated out.

The top-level joint deviance is

$$D_j\left(\mathbf{z}, \xi, \sigma^2, \gamma, \mathbf{T}\right) = -2 \log\left\{p\left(\mathbf{z}, \theta, \beta \left| \xi, \sigma^2, \gamma, \mathbf{T} \right.\right)\right\}, \tag{14}$$

and the corresponding likelihood is

$$p\left(\mathbf{z}, \theta, \beta \left| \xi, \sigma^2, \gamma, \mathbf{T} \right.\right) = p\left(\mathbf{z} \left| \xi, \theta \right.\right) p\left(\theta \left| \sigma^2, \beta \right.\right) p\left(\beta \left| \gamma, \mathbf{T} \right.\right). \tag{15}$$

The DIC calculated using the deviance in Equation 14 will be called the top-level joint DIC (DICj_T). It is a complete DIC, which contains all the information from the model.

As both marginal likelihood and joint likelihood are used frequently in statistical analysis/inference (Bjørnstad, 1996), DIC could be formed using either form of the likelihood. Celeux et al. (2006) provided a thorough discussion about the performances of different versions of DIC in a missing data model; however, their findings may not be easily generalizable to hierarchical models because the missing data (due to random effects) can occur at different levels of the model. Hence, our study will provide a unique contribution to the literature by comparing the performances of the variants of DIC under various conditions often seen in hierarchical models.

Table 1
Models Used in Simulation Studies

	Measurement Model	Individual-Level Model	School-Level Model
Model 1	2PNO	$\theta_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + e_{ij}$	$\beta_{0j} = \gamma_{00} + \gamma_{01}\omega_{1j} + u_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11}\omega_{1j} + u_{1j}$
Model 2	2PNO	$\theta_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + e_{ij}$	$\beta_{0j} = \gamma_{00} + u_{0j}$ $\beta_{1j} = \gamma_{10} + u_{1j}$
Model 3	2PNO	$\theta_{ij} = \beta_{0j} + e_{ij}$	$\beta_{0j} = \gamma_{00} + u_{0j}$
Model 4	1PNO	$\theta_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + e_{ij}$	$\beta_{0j} = \gamma_{00} + \gamma_{01}\omega_{1j} + u_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11}\omega_{1j} + u_{1j}$

Note. The notations are deferred to Equations 1 to 3.

In Equations B14 and B15, $\mathbf{H}_j$ is a Kn_j-by-Kn_j symmetry matrix. The inverse of $\mathbf{H}_j$ should be calculated $M \times J$ times when calculating the top-level marginalized DIC, where M is the number of interim values from post-burn-in iteration, which are used to calculate the posterior mean, and J is the number of groups (see Equation 1). When the number of interim values or groups is large, yielding a high-dimensional matrix, taking the matrix inverse becomes computationally prohibitive. Therefore, the top-level marginalized DIC is computationally much more demanding than other variants of DIC.

Simulation Studies

In this section, simulation studies are designed to evaluate the performance of the five versions of DIC in terms of selecting the correct model. The true multilevel IRT model differs by (1) whether significant individual- and/or school-level covariates are included; (2) whether 2PNO or 1PNO is used as the true measurement model. Two simulation studies are performed and they are described in detail below.

Simulation Design

For meaningful examination of the behavior of DICs, the simulated data must be generated from a plausible model (Sinharay & Stern, 2003).There are four MLIRT models to be chosen, named Models 1 to 4. Table 1 shows the specifications of the four models.

Tables 2 and 3 show the simulation design for Study 1 and Study 2, respectively. There are 36 (2 group size $\times$ 2 test length $\times$ 3 number of covariates $\times$ 3 measurement models) conditions in Study 1 and 8 (2 group size $\times$ 2 test length $\times$ 2 measurement models) conditions in Study 2. It is typical to consider one covariate in each level of structure models, because if more than one covariate in the higher-level of structure models (i.e., level 3 in this study) are considered, the 95% posterior credible interval (P.C.I.) can be calculated as a variable selection index, which we also report in Study 1. The Gibbs sampler is used to obtain the estimated values of the parameters. The source code is available to readers upon request.

Table 2
Fixed and Manipulated Conditions and Parameter Values in Study 1

Study 1	Generation Model	Model 1 or Model 2 or Model 3
(Fox & Glas, 2001)	Calibration model	Model 1 and Model 2 and Model 3
	Fixed conditions	
	Examinee sample size	2,000
	a_k Item discrimination	$a_k \sim \log N(\mu_a = \exp(1), \sigma_a = .25)$
	b_k Item difficulty	$b_k \sim N(\mu_b = 0, \sigma_b = .5)$
	x_{1j} Individual-specific covariate	$x_{1j} \sim N(\mu_x = 0, \sigma_x = 1)$
	e_{ij} Individual-specific random effect parameter	$e_{ij} \sim N(\mu_e = 0, \sigma_e = .2)$
	ω_{1j} School-specific covariate	$\omega_{1j} \sim N(\mu_\omega = .5, \sigma_\omega = 1)$
	γ School-specific regression coefficients	$\gamma_{00} = -.30, \gamma_{01} = .15$ $\gamma_{10} = .35, \gamma_{11} = 1.0$
	$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}$ School-specific random effect parameters	$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim$ MVN $\left(\mu_u = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_u = \begin{bmatrix} .1 & 0 \\ 0 & .1 \end{bmatrix}\right)$
	Manipulated conditions	
	Group size	Small: 10 Large: 200
	Test length	Small: 10 Large:30
	Number of covariates	One individual-specific covariate and one school-specific covariate One individual-specific covariate but no school-specific covariate No covariate

Without loss of generality, an additional simulation check was done based on Cohen, Kane, and Kim's (2001) index to detect the number of replications. Take θ as an example, let $r = 1, \ldots, R$ denote replications. Cohen et al. (2001) calculated a magnitude of the differences between the average of MSE(θ) (denoted as AMSE(θ)) under manipulated conditions. Then under a stringent tolerance criterion, we can obtain that

$$\hat{R} \geq \frac{\left(\sigma^2_{\text{MSE}(\theta)_{c1}} + \sigma^2_{\text{MSE}(\theta)_{c2}}\right) \Big/ R}{0.1 \times |\text{AMSE}(\theta)_{c1} - \text{AMSE}(\theta)_{c2}|}, \tag{16}$$

where the subscript "$c1$" and "$c2$" denote two different conditions, $\sigma^2_{\text{MSE}(\theta)}$ is the standard deviation of MSE(θ), and $\hat{R}$ denotes the least necessary number of replications.

Under the simulation conditions, when R was set to 50, the right side of Equation 16 was always less than 10, which means 10 or more replications are enough. Therefore, we consider 50 to be a reasonable and adequate number for this study.

Table 3
Fixed and Manipulated Conditions and Parameter Values in Study 2

Study 2	Generation Model	Model 1 or Model 4
	Calibration model	Model 1 and Model 4
	Fixed conditions	
	Examinee sample size	2,000
	Number of covariates	One individual-specific covariate and one school-specific covariate
	b_k,x_{1j},e_{ij},ω_{1j},γ,$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}$	The same as those in Study 1 (Table 2).
	Manipulated conditions	
	Group size	Small: 10
		Large: 200
	Test length	Small: 10
		Large: 30
	a_k Item discrimination	2PNO: $a_k \sim \log N(\mu_a = \exp(1), \sigma_a = .25)$
		1PNO: $a_k = 1$

Geweke's (1992) convergence diagnosis method was used to diagnose convergence. Three thousand iterations were treated as the initial phase; after that, under Geweke's approach, for a given parameter, a z score, which is defined as the difference between the first n_A ($n_A = 1{,}000$) and the last n_B ($n_B = 1{,}000$) iterations is computed as evidence of convergence,

$$z_\theta = \frac{\bar{\theta}^A - \bar{\theta}^B}{\sqrt{n_A^{-1}\hat{S}_\theta(0)^A - n_B^{-1}\hat{S}_\theta(0)^B}}, \tag{17}$$

where $\bar{\theta}$ denotes the sample mean of θ and $\hat{S}_\theta(0)$ denotes the consistent spectral density estimate. The z score tends to follow a standard normal distribution as $n \to \infty$ (Geweke, 1992). Hence, a z score less than 1.96 implied parameter convergence.

Through all the conditions, the Markov chain stabilized after 5,000 iterations. Hence, a chain length of 10,000 iterations with a burn-in of 5,000 is chosen reasonably for this study. We sampled one out of 20 points from the sampling phase to calculate the model selection criteria.

Result of Study 1

Table 4 presents the proportion of correct model selection for each of the five variants of DIC in Study 1. The values in the table indicate, out of the three fitted models, how often each of the indices selected the true model.

DIC_C, based on the measurement model, chose the correct model for both test lengths and both group sizes when Model 1 or 3 was the generation model. However, when the data were generated from Model 2, when the test length was small, DIC_C chose Model 1 44% of the time and Model 2 56% of the time for the small group size, and it chose Model 1 58% of the time and Model 2 42% of the time for the large group size, and when the test length was large, DIC_C chose Model 1 46% of the time

Table 4
Correct Model Selection Proportion in Study 1

			Generation Model														
			DIC_C			DIC_S			DIC_T			$DIC j_S$			$DIC j_T$		
K	J	Calibration Model	M 1	M 2	M 3	M 1	M 2	M 3	M 1	M 2	M 3	M 1	M 2	M 3	M 1	M 2	M 3
10	10	Model 1	**1.00**	.44	0	.02	.18	0	.02	.16	0	**1.00**	.18	0	0	0	0
		Model 2	0	**.56**	0	**.98**	**.82**	0	**.98**	**.84**	0	0	**.82**	0	**1.00**	.10	0
		Model 3	0	0	**1.00**	0	0	**1.00**	0	0	**1.00**	0	0	**1.00**	0	**.90**	**1.00**
	200	Model 1	**1.00**	**.58**	.04	**1.00**	**.50**	0	**1.00**	.28	0	**1.00**	**.54**	0	**.94**	.22	0
		Model 2	0	.42	.02	0	**.50**	0	0	**.36**	0	0	.46	0	0	.26	0
		Model 3	0	0	**.94**	0	0	**1.00**	0	**.36**	**1.00**	0	0	**1.00**	.06	**.54**	**1.00**
30	10	Model 1	**1.00**	.46	0	.08	.24	0	.16	.16	0	**.98**	.28	0	.32	.02	.18
		Model 2	0	**.54**	0	**.92**	**.76**	0	**.84**	**.84**	0	.02	**.72**	0	**.54**	.20	0
		Model 3	0	0	**1.00**	0	0	**1.00**	0	0	**1.00**	0	0	**1.00**	.14	**.78**	**.82**
	200	Model 1	**1.00**	**.54**	0	**1.00**	.42	0	**1.00**	.28	0	**1.00**	.48	0	**1.00**	.16	0
		Model 2	0	.46	0	0	**.58**	0	0	**.36**	0	0	**.52**	0	0	.12	0
		Model 3	0	0	**1.00**	0	0	**1.00**	0	**.36**	**1.00**	0	0	**1.00**	0	**.72**	**1.00**

Note. K = test length; J = group size; DIC_C = conditional DIC; DIC_S = second-level marginalized DIC; DIC_T = top-level marginalized DIC; $DIC j_S$ = second-level joint DIC; $DIC j_T$ = top-level joint DIC; M 1 = Model 1; M 2 = Model 2; M 3 = Model 3.

and Model 2 54% of the time for the small group size, and it chose Model 1 54% of the time and Model 2 46% of the time for the large group size. This observation indicates that DIC_C cannot easily distinguish models that differ by person-level covariates.

DIC_S, based on second-level marginal likelihood, chose the correct model for both test lengths and both group sizes when Model 3 was the true model. When Model 1 was the generation model for both test lengths, DIC_S could choose the correct model for the large group size; however, for the small group size, DIC_S tended to select Model 2 predominately (with the proportion of selection higher than .90). When the data were generated from Model 2, for small test length, DIC_S chose Model 1 18% of the time and Model 2 82% of the time for the small group size, and chose Model 1 and Model 2 with the same probability for the large group size, and for large test length, DIC_S chose Model 1 24% of the time and Model 2 76% of the time for the small group size, and chose Model 1 42% of the time and Model 2 58% of the time for the large group size. It appears from the results that Models 1 and 2 are relatively difficult to differentiate using DIC_S, and DIC_S tended to favor Model 2 when the group size is small.

DIC_T, based on the top-level marginal likelihood, chose the correct model when the true model was Model 3. When Model 1 was the true model, DIC_T chose Model 1 2% of the time and Model 2 98% of the time for the small test length and the small group size, for large test length, DIC_T chose Model 1 16% of the time and Model 2 84% of the time for the small group size, and for both test lengths and the large group size, it could choose the generation model with probability 1. When the data were generated from Model 2, the results for both test lengths were the same. DIC_T selected the true model 84% of the time for the small group size, and it chose Model 2 only approximately one-third of the time for the large group size. In other words, when Model 2 is the true model and when the group size is large, DIC_T cannot distinguish the three models very well.

Based on second-level joint likelihood, when the true model was Model 1 or Model 3, $DICj_S$ could choose the true model with probability higher than .98. When the data were generated from Model 2, when the test length was small, $DICj_S$ chose the true model 82% of the time for the small group size, and for the large group size it chose Model 1 54% of time and Model 2 46% of the time, and when the test length was large, $DICj_S$ chose the true model 72% of the time for the small group size, and for the large group size it chose Model 2 52% of the time. Overall, $DICj_S$ can almost select the true model for all manipulated conditions.

When Model 1 was the generation model, $DICj_T$, based on top-level joint likelihood, chose Model 2 with probability 1 for the small test length and the small group size, chose Model 1 as the best-fitting model 32% of the time and chose Model 2 54% of the time for the large test length and the small group size, and for the large group size it chose the correct model with probability larger than .94. When the true model was Model 2 or Model 3, $DICj_T$ chose Model 3 with probability higher than .09 for the small test length except when the test length was small and the group size was larger; under that condition, $DICj_T$ chose Model 3 with probability .54 when response data were generated from Model 2 and with probability 1 when data were generated from Model 3, and it chose Model 3 with probability higher than .72 for

Table 5
Proportion of the 95% P.C.I.s of $\boldsymbol{\gamma} = [\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}]$ *That Contain Zero in Study 1*

			Generation Model											
			Model 1				Model 2				Model 3			
K	*J*	Calibration Model	γ_{00}	γ_{01}	γ_{10}	γ_{11}	γ_{00}	γ_{01}	γ_{10}	γ_{11}	γ_{00}	γ_{01}	γ_{10}	γ_{11}
10	10	Model 1	.52	**.84**	.54	0	.44	**.98**	.46	**.96**	1	**1**	**1**	**1**
		Model 2	.68	–	0	–	.30	–	.26	–	1	–	**1**	–
		Model 3	.46	–	–	–	.20	–	–	–	1	–	–	–
	200	Model 1	0	0	0	0	0	**.86**	0	**.82**	0	**1**	**1**	**1**
		Model 2	0	–	0	–	0	–	0	–	0	–	**1**	–
		Model 3	1	–	–	–	0	–	–	–	0	–	–	–
30	10	Model 1	.60	**.84**	.42	0	.64	**.96**	.48	**.98**	1	**1**	**1**	**1**
		Model 2	.64	–	0	–	.06	-	0	–	1	–	**1**	–
		Model 3	.62	–	–	–	.06	–	–	–	0	–	–	–
	200	Model 1	0	0	0	0	0	**.94**	0	**.84**	0	**1**	**1**	**1**
		Model 2	.02	–	0	–	0	–	0	–	0	–	**1**	–
		Model 3	.02	–	–	–	0	–	–	–	0	–	–	–

Note. K = test length; J = group size. '-' means null.

the large test length. According to the results, Models 1 and 2 are relatively difficulty to differentiate using DICj_T, and DICj_T cannot easily distinguish models that differ by school-level covariates.

As shown in Table 4, the test length cannot affect the model selection by DICs. When Model 1 was the true model, for a small group size DIC$_C$ and DICj_S performed best, and for a large group size all variants of DIC performed well. When Model 2 was the true model, all variants of DIC had low sensitivities for the large group size, DIC$_S$ and DICj_S could choose the true model with probability higher than .50, and DIC$_C$ and DIC$_T$ performed similarly to DIC$_S$ and DICj_S when the group size was small. When Model 3 was the true model, all variants of DIC could choose the correct model with probability higher than .82.

Table 5 presents the proportion of the 95% P.C.I.s for school-specific regression coefficient $\boldsymbol{\gamma}$ which contain 0 in Study 1. We consider the proportion of the 95% P.C.I.s that contain 0 as an alternative model selection index to evaluate whether the inclusion of covariates is needed in the model. This is because Models 1 to 3 differ essentially on whether a certain covariate is included in the model. As shown in Table 5, the structural multilevel model of the generation model can be chosen with probability higher than .82 with the proportion of the 95% P.C.I.s that contain 0.

Result of Study 2

Table 6 presents the correct model selection proportion in Study 2. When the data were generated from Model 1, DIC$_C$ and DICj_S could choose the correct model with probability higher than .68, and the correct model could be selected by DIC$_C$, DICj_S,

Table 6
Correct Model Selection Frequency of Study 2

			Generation Model									
			DIC_C		DIC_S		DIC_T		$DICj_S$		$DICj_T$	
K	J	Calibration Model	M 1	M 4	M 1	M 4	M 1	M 4	M 1	M 4	M 1	M 4
10	10	Model 1	**.92**	.22	0	0	0	0	**.84**	.34	.42	.44
		Model 4	.08	**.78**	**1.00**	**1.00**	**1.00**	**1.00**	.16	**.66**	**.58**	**.56**
	200	Model 1	**.86**	.44	0	0	.04	.10	**.68**	.46	**.54**	**.50**
		Model 4	.14	**.56**	**1.00**	**1.00**	**.90**	**.90**	.32	**.54**	.46	**.50**
30	10	Model 1	**1.00**	.04	0	0	.0	0	**1.00**	.04	**1.00**	**.56**
		Model 4	0	**.96**	**1.00**	**1.00**	**1.00**	**1.00**	0	**.96**	0	.44
	200	Model 1	**1.00**	.12	0	0	.14	.24	**1.00**	.10	**1.00**	.42
		Model 4	0	**.88**	**1.00**	**1.00**	**.86**	**.76**	0	**.90**	0	**.58**

Note. K = test length; J = group size; DIC_C = conditional DIC; DIC_S = second-level marginalized DIC; DIC_T = top-level marginalized DIC; $DICj_S$ = second-level joint DIC; $DICj_T$ = top-level joint DIC; M 1 = Model 1; M 4 = Model 4.

and $DICj_T$ with probability 1 for the large test length. When Model 4 was the generation model, all indices except $DICj_T$ could choose the generation model with probability higher than .54, and $DICj_T$ could choose each model with similar probability. It appears that longer test length helps distinguish the number of item parameters in measurement models.

Real-Data Illustration

In this section, we present a real-data example to illustrate the application of the DIC indices. A data set from the Program for International Student Assessment (PISA) 2012 assessment (http://www.oecd.org/pisa/data/pisa2012databasedownloadabledata.htm) was analyzed.

Data Source

The PISA, collected by the Organization for Economic Co-operation and Development (OECD), is conducted to assess students' performance and explore the effects of student and institutional factors on student performance. In 2012, 65 countries participated in the assessment, and the survey covered mathematics, reading, science, and problem solving. In this section, we will focus only on the mathematics test given to 15-year-old United States students in 2012. The data set contains 4,978 students from 162 schools, and each student responded to 49 multiple choice items scored dichotomously. In addition, we chose the indices of economic, social, and cultural status (ESCS) and school location as individual-level and school-level covariates, respectively. For data cleaning, we deleted all schools in which there were fewer than 10 students. The resulting sample size entered into the final analysis was 4,882 students from 154 schools.

Table 7
Model Comparison Results for the PISA Example

	DIC_C(Rank)	DIC_S(Rank)	DIC_T(Rank)	$DICj_S$(Rank)	$DICj_T$(Rank)
Model 1	91166.62(1)	52,9868.71(5)	52,9904.25(5)	89,426.13(2)	11,4476.65(4)
Model 2	91,167.25(2)	52,9937.53(6)	52,9958.12(6)	89,450.28(3)	11,4275.47(3)
Model 3	91,563.14(3)	52,8752.41(4)	52,8709.88(4)	88,273.01(1)	78,486.09(1)
Model 4	93,343.43(4)	52,3447.23(2)	52,3607.71(2)	93,913.58(5)	12,0908.47(5)
Model 5	93,348.21(5)	52,3454.88(3)	52,3618.39(3)	93,913.65(6)	12,0996.91(6)
Model 6	93,977.81(6)	52,2507.45(1)	52,2553.78(1)	93,507.10(4)	83,857.69(2)

Note. DIC_C = conditional DIC; DIC_S = second-level marginalized DIC; DIC_T = top-level marginalized DIC; $DICj_S$ = second-level joint DIC; $DICj_T$ = top-level joint DIC.

Table 8
95% P.C.I.s of γ *for the PISA Example*

		Est.	P.C.I.			Est.	P.C.I.
Model 1	γ_{00}	−.0358	[−.0959,.0239]	Model 4	γ_{00}	−.0291	[−.0905,.0341]
	γ_{01}	−.00018	[−.0192,.0160]		γ_{01}	−.0039	[−.0222,.0147]
	γ_{10}	.0192	[−.0420,.0792]		γ_{10}	.0474	[−.0209,.1129]
	γ_{11}	−.0019	[−.0196,.0163]		γ_{11}	.0031	[−.0217,.0165]
Model 2	γ_{00}	−.0416	[−.0630,−.0200]	Model 5	γ_{00}	−.0416	[−.0637,−.0188]
	γ_{10}	.0131	[−.0081,.0344]		γ_{10}	.0368	[.0145,.0599]
Model 3	γ_{00}	−.037	[−.0592,−.0156]	Model 6	γ_{00}	−.0323	[−.0548,−.0107]

Estimation of Model Parameters

Six models were applied to the PISA data; they were Models 1 to 4, described in the simulation studies, in which ESCS was considered as the individual-level covariate, and school location was considered as the school-level covariate, along with two additional models: the 1PNO model as the measurement model, and a structural multilevel model similar to Model 2 and Model 3. For the MCMC algorithm, according to Geweke's convergence criterion, a conservative burn-in of 5,000 iterations and 5,000 post-burn-in iterations was used here.

Results

Model selection results for the real-data illustration are given in Tables 7 and 8. As noted above, the smaller the DIC, the better the model–data fit. Meanwhile, a popular rule of thumb for model comparison (e.g., Spiegelhalter et al., 2002), is that a difference of 2 or less is considered negligible, a difference between 3 and 7 provides positive support for the model with a lower value, and a difference exceeding 7 constitutes strong support.

As shown in Table 7, because the difference in the DIC_C values between Model 1 and Model 2 was less than 1, the two models are nearly equally favorable based on DIC_C. Model 3 was selected by $DICj_S$, and $DICj_T$ and Model 6 was selected by DIC_S

and $DIC_{T,}$ because they generated the smallest DIC value. Table 8 provides the 95% P.C.I.s for the coefficient γ. As seen in Table 8, when comparing Model 1 to Model 3, it appears that the 95% P.C.I.s of γ contained 0 in Model 1 and Model 2, indicating that these parameters are not significantly different from 0. Conversely, the P.C.I.s of γ_{00} in Model 3 did not contain 0, implying that this parameter must be included in the model. Taken together, Model 3 is the preferred model among the three fitted models, of which 2PNO models are the measurement models. When comparing Model 4 to Model 6, this observation indicates that the 95% P.C.I.s of γ contained 0 in Model 4, indicating that these parameters are not significantly different from 0. Conversely, the P.C.I.s of γ_{00} in Models 5 and 6 and the P.C.I.s of γ_{10} in Model 5 did not contain 0, implying that those parameters must be included in the model. Taken together, Model 5 is the preferred model among the three fitted models of which the 1PNO models are the measurement models. Considering the results from all indices, it appeared that Model 3 may be the best fit for the data.

Discussion

The present study was motivated by two observations. First, there is a lack of effective model selection criteria for multilevel IRT models. Because the MCMC sampling method is often used to estimate multilevel IRT models, DIC becomes a natural option for evaluating the global model fit. Second, for multilevel models, DIC can be constructed differently depending on (1) whether joint or marginal likelihood is used and (2) the target parameters of interest. There is currently not enough information on which version of DIC is recommended for multilevel IRT models when different measurement models are considered and when various levels of covariates are included in the model. The main purpose of this study was to examine the accuracy of five DIC-based indices in the selection of a best-fitting MLIRT model.

Across all simulation conditions, second-level joint DIC is recommended for MLIRT models because it almost selects the correct model and is computationally less demanding than some other alternatives. Conditional DIC sometimes tends to choose a more complex model for a large group size. Second-level marginalized DIC and top-level marginalized DIC perform similarly; they tend to choose a simpler model for a small group size. For a large group size, top-level marginalized DIC performs worse than second-level marginalized DIC. Because the numerical complexity of top-level marginalized DIC is far more demanding, top-level marginalized DIC is not recommended for all conditions for multilevel models, and top-level joint DIC tend to select a simpler model regardless of the group size.

Additionally, we conducted a sensitivity analysis and found that the computed joint DICs showed more variation than the marginalized DICs. This happened because, when to calculating the joint DICs, the point estimates of individual θ and ξ were plugged in and these point estimates are prone to measurement errors. However, when calculating the marginalized DICs, the influences of random effects were integrated out in the marginal likelihood. That is the reason why the joint DICs have more variation, especially the top-level joint DIC.

Because this study was a preliminary investigation of the comparison of joint DIC and marginalized DIC for MLIRT models, there are several limitations. First, from

the empirical perspective, only one individual-specific covariate and one school-specific covariate were considered in the real-data illustration; future research should assess the effect of more covariates at both the individual level and school level. Second, the performance of DIC-based indices in this study was focused on a multilevel data structure at only one time point. Future research surrounding the model selection method for MLIRT models can be expanded to select multilevel models for longitudinal data. Finally, as warned by Celeux et al. (2006) and Li et al. (2009), DIC was less accurate with mixture models; one might wish to investigate the restrictions under which some type of DIC can perform well for multilevel mixture models.

As a point of reference, it should be noted that the computation of top-level marginalized DIC is extremely time-consuming. Several days of CPU time was required on a 3.20 GHz desktop PC to compute per condition, per replication using MATLAB 2013a. Other versions of DIC required less than 10 minutes.

Acknowledgments

This research was supported by Key Laboratory of Applied Statistics of MOE, National Natural Science and Social Science Foundations of China (Grants 11571069) and Institute of Education Sciences (IES) Grant R305D160010. We would like to thank the Editor in Chief Dr. George Engelhard Jr. and Dr. Jonathan Templin, the associate editor, and three anonymous reviewers for their helpful comments on the earlier draft of this article. We are also thankful to Dr. Jiwei Zhang for his input.

Appendix 1: MCMC Algorithm for the MLIRT Model

In this appendix, the MCMC algorithm used in this article is described briefly. Interested readers can refer to Fox (2010) for a full description of the MCMC algorithm for the family of MLIRT models.

The full posterior distribution of the parameters given the data is given by

$$p(\mathbf{z}, \theta, \xi, \beta, \sigma^2, \gamma, \mathbf{T} \,|\mathbf{Y}, \mathbf{X}, \mathbf{W}) \propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} \left(\prod_{k=1}^{K} p(z_{ijk} \,|\theta_{ij}, \xi_k, y_{ijk}) \right) p(\theta_{ij} \,|\beta_j, \sigma^2, \mathbf{X}_j) \cdot p(\beta_j \,|\gamma, \mathbf{T}, \mathbf{W}_j) p(\gamma \,|\mathbf{T}) p(\xi) p(\sigma^2) p(\mathbf{T}) \quad \text{(A1)}$$

Step 1: Sampling $\mathbf{z}$. Given the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, the variables z_{ijk} are independent, according to the definition of z_{ijk}, it follows that

$$p(z_{ijk} \,|\theta_{ij}, \xi_k, y_{ijk}) \propto \psi(z_{ijk}; a_k\theta_{ij} - b_k, 1)[I(z_{ijk} > 0)I(y_{ijk} = 1) + I(z_{ijk} \leq 0)I(y_{ijk} = 0)], \quad \text{(A2)}$$

where $\psi(\cdot; a_k\theta_{ij} - b_k, 1)$ denotes the normal density with a mean equal to $a_k\theta_{ij}$ - b_k and a variance equals to one, and $I(\cdot)$ is an indicator variable which equals to one if its argument is true, and equals to zero otherwise.

So the fully conditional posterior density of z_{ijk} is given by

$$z_{ijk} \,\Big|\, \theta, \xi, \mathbf{Y} \sim \begin{cases} N(a_k\theta_{ij} - b_k, 1) \text{ truncated at the left by 0 if } y_{ijk} = 1 \\ N(a_k\theta_{ij} - b_k, 1) \text{ truncated at the right by 0 if } y_{ijk} = 0 \end{cases}. \quad \text{(A3)}$$

Step 2: Sampling $\boldsymbol{\theta}$. The ability parameters are independent given $\mathbf{z}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}$ and σ^2. Its full conditional posterior density is given by

$$\theta_{ij} \left| \mathbf{z}_{ij}, \xi, \beta_j, \sigma^2 \sim N \left(\frac{\hat{\theta}_{ij}/v + \mathbf{x}_{ij}\beta_j/\sigma^2}{1/v + 1/\sigma^2}, \frac{1}{1/v + 1/\sigma^2} \right) , \right. \tag{A4}$$

with

$$\hat{\theta}_{ij} = \frac{\sum_{k=1}^{K} a_k(z_{ijk} + b_k)}{\sum_{k=1}^{K} a_k^2}, \quad v = \left(\sum\nolimits_{k=1}^{K} a_k^2 \right)^{-1}. \tag{A5}$$

Step 3: Sampling $\boldsymbol{\xi}$. Let $\mathbf{E} = [\theta, -1]$, therefore

$$\xi_k \left| \theta, \mathbf{Z}_k \sim N(\hat{\xi}_k, (\mathbf{E}^t\mathbf{E})^{-1}) I(a_k > 0) \right. , \tag{A6}$$

with $\hat{\xi}_k = (\mathbf{E}^t\mathbf{E})^{-1}\mathbf{E}\mathbf{Z}_k$. Hereafter, the superscript t denotes the transposition of a matrix.

Step 4: For each j, sample $\boldsymbol{\beta}_\mathrm{j}$ from the full conditional

$$\beta_j \left| \theta_j, \sigma^2, \gamma, \mathbf{T} \sim N(\mathbf{D}\mathbf{d}, \mathbf{D}) \right. , \tag{A7}$$

where $\mathbf{D} = (\Sigma_j^{-1} + \mathbf{T}^{-1})^{-1}$ and $\mathbf{d} = \Sigma_j^{-1}\hat{\beta}_j + \mathbf{T}^{-1}\omega_j\gamma$, with $\Sigma_j = \sigma^2(\mathbf{x}_j^t\mathbf{x}_j)^{-1}$ and $\hat{\beta}_j = (\mathbf{x}_j^t\mathbf{x}_j)^{-1}\mathbf{x}_j^t\theta_j$.

Step 5: Sample $\boldsymbol{\gamma}$ from the full conditional

$$\gamma \left| \beta_j, \mathbf{T} \sim N \left(\left(\sum\nolimits_{j=1}^{J} \omega_j^t\mathbf{T}^{-1}\omega_j \right)^{-1} \sum\nolimits_{j=1}^{J} \omega_j^t\mathbf{T}^{-1}\beta_j, \left(\sum\nolimits_{j=1}^{J} \omega_j^t\mathbf{T}^{-1}\omega_j \right)^{-1} \right) . \right. \tag{A8}$$

Step 6: Sample σ^2 given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ from the full conditional

$$\sigma^2 \left| \theta, \beta \sim Inv - \chi^2(N, S^2), \right. \tag{A9}$$

where $S^2 = \frac{1}{N}\sum_{j=1}^{J}(\theta_j - \mathbf{x}_j\beta_j)^t(\theta_j - \mathbf{x}_j\beta_j)$.

Step 7: Sample $\mathbf{T}$ from the full conditional

$$\mathbf{T} \left| \beta, \gamma \sim Inv - Wishart(J, \mathbf{S}_T^{-1}), \right. \tag{A10}$$

where $\mathbf{S}_T^{-1} = \sum_{j=1}^{J}(\beta_j - \omega_j\gamma)(\beta_j - \omega_j\gamma)^t$.

Appendix 2: Calculation of Variants of DIC

The First-Level Conditional DIC

The density function of $\mathbf{z}$ given $(\boldsymbol{\xi}, \boldsymbol{\theta})$ in Equation 7 can be expressed as

$$p(\mathbf{z} \left| \xi, \theta \right.) = \prod_j (2\pi)^{-Kn_j/2} \exp \left\{ -\frac{1}{2} \sum_{i,k} \left(z_{ijk} + b_k - a_k\theta_{ij} \right)^2 \right\}, \tag{B1}$$

where n_j is the number of examinees in group j. θ_{ij} is normally distributed with variance $\Omega_\theta = (\mathbf{a}^t\mathbf{a} + \sigma^{-2})^{-1}$ and mean $\mu_\theta = \Omega_\theta(\mathbf{a}^t(\mathbf{z}_{ij} + \mathbf{b}) + \sigma^{-2}\mathbf{x}_{ij}^t\beta_j)$, $\mathbf{a}$ is the vector of discrimination parameters, $\mathbf{b}$ is the vector of difficulty parameters, and $\mathbf{z}_{ij}$ and $\mathbf{x}_{ij}$ are the vectors of the augmented data and the individual-specific covariates of examinee i in group j, respectively.

Hence, the augmented likelihood, focused on the parameters from level 1, used to calculate the posterior mean of the deviance is expressed as

$$p\left(\mathbf{z}^{(r)}\left|\xi,\theta^{(r)}\right.\right)=\prod_{j}(2\pi)^{-Kn_j/2}\exp\left\{-\frac{1}{2}\sum_{i,k}\left(z_{ijk}^{(r)}+b_k^{(r)}-a_k^{(r)}\theta_{ij}^{(r)}\right)^2\right\}. \quad \text{(B2)}$$

Hereafter, the superscript r denotes the interim value from the rth post-burn-in iteration.

In contrast, the deviance evaluated at the posterior mean $\hat{\theta}$ can be expressed as

$$p\left(\hat{\mathbf{z}}\left|\xi,\hat{\theta}\right.\right)=\prod_{j}(2\pi)^{-Kn_j/2}\exp\left\{-\frac{1}{2}\sum_{i,k}\left(\hat{z}_{ijk}+\hat{b}_k-\hat{a}_k\hat{\theta}_{ij}\right)^2\right\}. \quad \text{(B3)}$$

Hereafter, the hat denotes the final point estimate value based on the Gibbs sampler, and $\hat{\mathbf{z}}$ denotes the augmented data based on $\mathbf{a}$, $\mathbf{b}$, and $\hat{\theta}$.

The Second-Level Marginalized DIC

Let $\boldsymbol{\Omega}=(\boldsymbol{\xi},\sigma^2,\boldsymbol{\beta})$ denote all model parameters of interest for DIC_S, where $\boldsymbol{\xi}$ denotes the matrix of item parameters, σ^2 denotes the variance of level 2 random effects (see Equation 2), and $\boldsymbol{\beta}$ denotes the matrix of regression coefficients in level 2. The method to calculate the second-level marginalized DIC is similar to Fox (2010). Interest is focused on the augmented likelihood in Equation 9,

$$p(\mathbf{z}\,|\xi,\,\sigma^2,\beta)\triangleq p(\mathbf{z}\,|\Omega)=\frac{p(\mathbf{z},\theta\,|\Omega)}{p(\theta\,|\mathbf{z},\,\Omega)}=\frac{p(\mathbf{z}\,|\xi,\theta)p(\theta\,|\sigma^2,\beta)}{p(\theta\,|\mathbf{z},\,\Omega)}, \quad \text{(B4)}$$

where $\mathrm{p}(\boldsymbol{\theta}|\mathbf{z},\boldsymbol{\Omega})$ is the full conditional probability density function, the density function of $\mathbf{z}$ given $(\boldsymbol{\xi},\boldsymbol{\theta})$ is the same as that in Equation B1, and the density function of $\boldsymbol{\theta}$ given $(\sigma^2,\boldsymbol{\beta})$ can be expressed as

$$p\left(\theta\left|\sigma^2,\beta\right.\right)=\prod_{j}(2\pi\sigma^2)^{-Kn_j/2}\exp\left[-\frac{\left(\theta_j-\mathbf{x}_j^t\beta_j\right)^t\left(\theta_j-\mathbf{x}_j^t\beta_j\right)}{2\sigma^2}\right]. \quad \text{(B5)}$$

Hence, the augmented likelihood used to calculate the posterior mean of the deviance follows that

$$p\left(\mathbf{z}_j^{(r)}\left|\Omega^{(r)}\right.\right)=(2\pi)^{-Kn_j/2}\left(\frac{\Omega_\theta^{(r)}}{\sigma^{2(r)}}\right)^{n_j/2}\exp\left\{-S^{(r)}\left(\theta_j^{(r)}\right)\Big/2\right\}, \quad \text{(B6)}$$

where

$$S^{(r)}\left(\theta_j^{(r)}\right)=\sum_{i,k}\left(z_{ijk}^{(r)}+b_k^{(r)}-a_k^{(r)}\theta_{ij}^{(r)}\right)^2+\sigma^{-2(r)}\left(\theta_j^{(r)}-\mathbf{x}_j\beta_j^{(r)}\right)^t\left(\theta_j^{(r)}-\mathbf{x}_j\beta_j^{(r)}\right). \quad \text{(B7)}$$

However, the deviance evaluated at the posterior mean $\hat{\Omega}$ follows that

$$p\left(\hat{\mathbf{z}}_j\left|\hat{\Omega}\right.\right)=(2\pi)^{-Kn_j/2}\left(\frac{\hat{\Omega}_\theta}{\hat{\sigma}^2}\right)^{n_j/2}\exp\left\{-\hat{S}\left(\hat{\theta}_j\right)\Big/2\right\}, \quad \text{(B8)}$$

where

$$\hat{S}\left(\hat{\theta}_j\right) = \sum_{i,k}\left(\hat{z}_{ijk} + \hat{b}_k - \hat{a}_k\hat{\theta}_{ij}\right)^2 + \hat{\sigma}^{-2}\left(\hat{\theta}_j - \mathbf{x}_j\hat{\beta}_j\right)^t\left(\hat{\theta}_j - \mathbf{x}_j\hat{\beta}_j\right). \tag{B9}$$

The Second-Level Joint DIC

The joint likelihood of all parameters of interest from levels 1 and 2 is written as

$$p\left(\mathbf{z}, \theta \,|\Omega\right) = \prod_{i,j,k}\psi\left(z_{ijk}^{(r)}; a_k^{(r)}\theta_{ij}^{(r)} - b_k^{(r)}, 1\right)\prod_{i,j}\psi\left(\theta_{ij}^{(r)}; \mathbf{x}_{ij}\beta_{ij}^{(r)}, \sigma^{2(r)}\right), \tag{B10}$$

where $\psi(\cdot\,; \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2. Moreover, the deviance evaluated at the posterior mean $\hat{\Omega}$ follows that

$$p(\hat{\mathbf{z}}, \hat{\theta}\,|\hat{\Omega}) = \prod_{i,j,k}\psi\left(\hat{z}_{ijk}; \hat{a}_k\hat{\theta}_{ij} - \hat{b}_k, 1\right)\prod_{i,j}\psi\left(\hat{\theta}_{ij}; \mathbf{x}_{ij}\hat{\beta}_{ij}, \hat{\sigma}^2\right). \tag{B11}$$

The Top-Level Marginalized DIC

Let $\mathbf{\Lambda} = (\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T})$, where $\mathbf{T}$ denotes the covariance of level 3 random effects (see Equation 3), and $\boldsymbol{\gamma}$ denotes the vector of regression coefficients in level 2. The method to calculate the top-level marginalized DIC refers to Fox's method (2010). Interested readers can refer to Fox (2010, pp. 190–191) for a detailed description.

Based on Bayes's formula, the augmented likelihood in Equation 13 can be obtained as follows:

$$p(\mathbf{z}\,|\xi, \sigma^2, \gamma, \mathbf{T}) \stackrel{\Delta}{=} p(\mathbf{z}\,|\Lambda) = p(\mathbf{z}\,|\beta, \Lambda)\frac{p(\beta\,|\gamma, \mathbf{T})}{p(\beta\,|\mathbf{z}, \Lambda)}. \tag{B12}$$

The first part of Equation B12, the conditional augmented likelihood given $(\boldsymbol{\beta}, \mathbf{\Lambda})$, is derived as

$$\begin{aligned} p\left(\mathbf{z}\,|\beta, \Lambda\right) = \prod_j (2\pi)^{-Kn_j/2}\left(\tfrac{\Omega_\theta}{\sigma^2}\right)^{n_j/2}\exp\left\{-\tfrac{1}{2}\left(z_{ijk} + b_k - a_k\theta_{ij}\right)^2\right\}, \\ \exp\left\{-\tfrac{1}{2\sigma^2}\left(\theta_j - \mathbf{x}_j\beta_j\right)^t\left(\theta_j - \mathbf{x}_j\beta_j\right)\right\}, \end{aligned} \tag{B13}$$

where θ_{ij} is normally distributed as in Equation B1.

Then, the density function of $\boldsymbol{\beta}_j$ given $(\boldsymbol{\gamma}, \mathbf{T})$ is a multivariate normal probability density function with mean $\boldsymbol{\omega}_j\,\boldsymbol{\gamma}$ and covariance matrix $\mathbf{T}$. The conditional distribution of $\boldsymbol{\beta}_j\,|\mathbf{z}_j, \mathbf{\Lambda}$ is multivariate normal with mean

$$E(\beta_j\,|\mathbf{z}_j, \Lambda) = \omega_j\gamma + (\mathbf{T}\mathbf{x}_j^t \otimes \mathbf{a}^t)\mathbf{H}_j^{-1}(\mathbf{z}_j - (\mathbf{x}_j\omega_j\gamma \otimes \mathbf{a} - \mathbf{1}_{n_j} \otimes \mathbf{b})), \tag{B14}$$

and covariance matrix

$$\Sigma_{\beta_j} = Var(\beta_j\,|\mathbf{z}_j, \Lambda) = \mathbf{T} + (\mathbf{T}\mathbf{x}_j^t \otimes \mathbf{a}^t)\mathbf{H}_j^{-1}(\mathbf{x}_j\mathbf{T} \otimes \mathbf{a}), \tag{B15}$$

where $\mathbf{H}_j^{-1} = \mathbf{x}_j^y\mathbf{T}\mathbf{x}_j \otimes \mathbf{a}\mathbf{a}^t + \mathbf{I}_{n_j} \otimes (\sigma^2\mathbf{a}\mathbf{a}^t + \mathbf{I}_K)$.

The deviance evaluated using the posterior mean, obtained by performing (B12), is given by

$$p\left(\mathbf{z}_j^{(r)}\,|\Lambda^{(r)}\right) = (2\pi)^{-Kn_j/2}\left(\frac{\Omega_\theta^{(r)}}{\sigma^{2(r)}}\right)^{n_j/2}\left|\mathbf{T}^{(r)}\right|^{-1/2}\left|\Sigma_\beta^{(r)}\right|^{1/2}\exp\left\{-B^{(r)}\left(\theta_j^{(r)}, \beta_j^{(r)}\right)/2\right\}, \tag{B16}$$

where

$$B^{(r)}\left(\theta_j^{(r)}, \beta_j^{(r)}\right) = \sum_{i,k}\left(z_{ijk}^{(r)} + b_k^{(r)} - a_k^{(r)}\theta_{ij}^{(r)}\right)^2 + \sigma^{-2(r)}\left(\theta_j^{(r)} - \mathbf{x}_j\beta_j^{(r)}\right)^t\left(\theta_j^{(r)} - \mathbf{x}_j\beta_j^{(r)}\right) + \left(\beta_j^{(r)} - \omega_j\gamma_j^{(r)}\right)^t\left(\beta_j^{(r)} - \omega_j\gamma_j^{(r)}\right). \quad \text{(B17)}$$

Moreover, the deviance evaluated at the posterior mean $\hat{\Lambda}$ follows that

$$p\left(\hat{\mathbf{z}}_j \,|\, \hat{\Lambda}\right) = (2\pi)^{-Kn_j/2}\left(\frac{\hat{\Omega}_\theta}{\hat{\sigma}^2}\right)^{n_j/2}\left|\hat{\mathbf{T}}\right|^{-1/2}\left|\hat{\Sigma}_\beta\right|^{1/2}\exp\left\{-\hat{B}\left(\hat{\theta}_j, \hat{\beta}_j\right)/2\right\}, \quad \text{(B18)}$$

where

$$\hat{B}\left(\hat{\theta}_j, \hat{\beta}_j\right) = \sum_{i,k}\left(\hat{z}_{ijk} + \hat{b}_k - \hat{a}_k\hat{\theta}_{ij}\right)^2 + \hat{\sigma}^{-2}\left(\hat{\theta}_j - \mathbf{x}_j\hat{\beta}_j\right)^t\left(\hat{\theta}_j - \mathbf{x}_j\hat{\beta}_j\right) + \left(\hat{\beta}_j - \omega_j\hat{\gamma}_j\right)^t\left(\hat{\beta}_j - \omega_j\hat{\gamma}_j\right). \quad \text{(B19)}$$

The Top-Level Joint DIC

The top-level joint likelihood used to calculate the posterior mean of the deviance in Equation 15 is then

$$p\left(\mathbf{z}^{(r)}, \theta^{(r)}, \beta^{(r)} \,|\, \Lambda^{(r)}\right) = \prod_{i,j,k}\psi\left(z_{ijk}^{(r)}; a_k^{(r)}\theta_{ij}^{(r)} - b_k^{(r)}, 1\right) \prod_{i,j}\psi\left(\theta_{ij}^{(r)}; \mathbf{x}_{ij}\beta_{ij}^{(r)}, \sigma^{2(r)}\right)\psi_{MVN}\left(\beta_{ij}^{(r)}; \omega_j\gamma^{(r)}, \mathbf{T}^{(r)}\right), \quad \text{(B20)}$$

where $\psi_{MVN}(\cdot\,; \mu, \Sigma)$ denotes the multivariate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Moreover, the deviance evaluated at the posterior mean $\hat{\Lambda}$ follows that

$$p\left(\hat{\mathbf{z}}, \hat{\theta}, \hat{\beta} \,|\, \hat{\Lambda}\right) = \prod_{i,j,k}\psi\left(\hat{z}_{ijk}; \hat{a}_k\hat{\theta}_{ij} - \hat{b}_k, 1\right)\prod_{i,j}\psi\left(\hat{\theta}_{ij}; \mathbf{x}_{ij}\hat{\beta}_{ij}, \hat{\sigma}^2\right)\psi_{MVN}\left(\hat{\beta}_{ij}; \omega_j\hat{\gamma}, \hat{\mathbf{T}}\right). \quad \text{(B21)}$$

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*(3), 251–269.

Ando, T. (2010). *Bayesian model selection and statistical modeling*. Boca Raton, FL: Chapman & Hall/CRC Press.

Ando, T., & Tsay, R. (2010). Predictive likelihood for Bayesian model selection and averaging. *International Journal of Forecasting*, *26*, 744–763.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Barnett, A. G., Koper, N., Dobson, A. J., Schmiegelow, F., & Manseau, M. (2010). Using information criteria to select the correct variance–covariance structure for longitudinal data in ecology. *Methods in Ecology and Evolution*, *1*(1), 15–24.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–335.

Bjørnstad, J. F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, *91*, 791–806.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution*, *24*(3), 127–135.

Carlin, B. P. (2006). Comment on article by Celeux et al. *Bayesian Analysis*, *1*, 675–676.

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*, 651–673.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.

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

Choi, I. H., & Wilson, M. (2016). Incorporating mobility in growth modeling for multilevel and longitudinal item response data. *Multivariate Behavioral Research*, *51*(1), 120–137.

Cohen, A. S., Kane, M. T., & Kim, S. H. (2001). The precision of simulation study results. *Applied Psychological Measurement*, *25*(2), 136–145.

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

DeMars, C. (2010). *Item response theory*. Oxford, UK: Oxford University Press.

Eklund, J., & Karlsson, S. (2007). Forecast combination and model averaging using predictive measures. *Econometric Reviews*, *26*, 329–363.

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

Entink, R. K., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48.

Fox, J.-P. (2010). *Bayesian item response modeling theory and applications*. New York, NY: Springer.

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

Geerlings, H., Glas, C. A., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, *76*, 337–359.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*, 153–60.

Gelfand, A. E. (1996). *Markov chain Monte Carlo in practice*. London, UK: Chapman & Hall.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B (Methodological)*, *56*, 501–514.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1996). *Bayesian data analysis*. London, UK: Chapman and Hall.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, A. F. M. Smith, A. P. Dawid, & J. O. Berger (Eds.), *Bayesian statistics* (pp. 169–193). New York, NY: Oxford University Press.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, UK: Anorld.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society, Series B (Methodological)*, *29*, 83–100.

Hamaker, E. L., van Hattum, P., Kuiper, R. M., & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook for Advanced Multilevel Analysis* (pp. 231–255). New York, NY: Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer.

Hung, L. F., & Wang, W. C. (2012). The generalized multilevel facets model for longitudinal data. *Journal of Educational and Behavioral Statistics*, *37*(2), 231–255.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon Press.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*, 331–358.

Kass, R., & Raftery, A. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association*, *90*, 773–795.

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, *90*, 928–934.

Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *American Statistician*, *53*(2), 119–122.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Meng, X. L., & Vaida, F. (2006). Comment on article by Celeux et al. *Bayesian Analysis*, *1*, 687–698.

Millar, R. B. (2009). Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factor. *Biometrics*, *65*, 962–969.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Plummer, M. (2006). Comment on Article by Celeux et al. *Bayesian Analysis*, *1*, 681–686.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., & Lewis, T. (2000). *A user's guide to MLwiN, Version 2.1*. London, UK: Institute of Education, University of London.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sinharay, S., & Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, *111*, 209–221.

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

Spiegelhalter, D., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS user manual*. Cambridge, UK: MRC Biostatistics Unit.

Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 144–168.

Authors

XUE ZHANG is a post-doc at China Institute of Rural Education Development, Northeast Normal University, 5268 Renmin Street, Changchun, 130024, Jilin, China; Zhangx815@nenu.edu.cn. Her primary research interests include model/item fit, Bayesian inference, longitudinal data analysis and item response theory.

JIAN TAO is a Professor at School of Mathematics and Statistics, Northeast Normal University, 5268 Renmin Street, Changchun, 130024, Jilin, China; taoj@nenu.edu.cn. His primary research interests include statistical methods, Bayesian inference, and item response theory.

CHUN WANG is an Assistant Professor at the University of Washington, Seattle, WA 98105, USA; wang4066@uw.edu. Her primary research interests include multidimensional/multilevel item response theory models, computerized adaptive testing, and cognitive diagnostics models.

NING-ZHONG SHI is a Professor at School of Mathematics and Statistics, Northeast Normal University, 5268 Renmin Street, Changchun, 130024, Jilin, China; shinz@nenu.edu.cn. His primary research interests include order restricted statistical inference and hypothesis testing.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table A1. Parameter Recovery in Study 1.
Table A2. Estimation of γ in Study 1.
Table A3. Parameter Recovery in Study 2.
Table A4. Summary of Item Parameter Estimates.
Figure A1. Empirical PDF plot of the θ estimates.