Model Selection for Multilevel Mixture Rasch Models

Applied Psychological Measurement
2019, Vol. 43(4) 272–289
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618779990
journals.sagepub.com/home/apm

Sedat Sen¹ , Allan S. Cohen² and Seock-Ho Kim²

Abstract

Mixture item response theory (MixIRT) models can sometimes be used to model the heterogeneity among the individuals from different subpopulations, but these models do not account for the multilevel structure that is common in educational and psychological data. Multilevel extensions of the MixIRT models have been proposed to address this shortcoming. Successful applications of multilevel MixIRT models depend in part on detection of the best fitting model. In this study, performance of information indices, Akaike information criterion (AIC), Bayesian information criterion (BIC), consistent Akaike information criterion (CAIC), and sample-size adjusted Bayesian information criterion (SABIC), were compared for use in model selection with a two-level mixture Rasch model in the context of a real data example and a simulation study. Level 1 consisted of students and Level 2 consisted of schools. The performances of the model selection criteria under different sample sizes were investigated in a simulation study. Total sample size (number of students) and Level 2 sample size (number of schools) were studied for calculation of information criterion indices to examine the performance of these fit indices. Simulation study results indicated that CAIC and BIC performed better than the other indices at detection of the true (i.e., generating) model. Furthermore, information indices based on total sample size yielded more accurate detections than indices at Level 2.

Keywords

mixture IRT, multilevel mixture Rasch model, model selection

Introduction

Standard unidimensional item response theory (IRT) models assume that a single model applies to all examinees in the population (Lord & Novick, 1968). However, examinees putatively from the same population may have response patterns which differ substantially enough to suggest that they come from different populations. Mixture IRT (MixIRT) models have been proposed as one way to handle this heterogeneity (Mislevy & Verhelst, 1990; Rost, 1990). Unlike standard IRT models, MixIRT models assume each person belongs to one of a fixed number of discrete latent populations referred to as latent groups or latent classes.

The MixIRT model is based on a combination of two different models, a latent class model and an IRT model. This synthesis makes it potentially possible to obtain qualitative and quantitative information simultaneously. Flexibility of MixIRT models with regard to examinee characteristics make these models useful tools for investigation of several psychometric issues, such

¹Harran University, Sanliurfa, Turkey

²University of Georgia, Athens, USA

Corresponding Author:

Sedat Sen, Faculty of Education, Harran University, Sanliurfa 63300, Turkey.

Email: sedatsen06@gmail.com

as detection of differential item functioning (DIF; Cohen & Bolt, 2005; Samuelsen, 2008), different response strategies (Bolt, Cohen, & Wollack, 2001; Mislevy & Verhelst, 1990), effects of testing accommodations (Cohen, Gregg, & Deng, 2005), and test speededness (Bolt, Cohen, & Wollack, 2002; Wang & Cohen, 2008).

Multilevel extensions of mixture IRT models (MMixIRT; Cho & Cohen, 2010; Vermunt, 2007) have been proposed to account for the multilevel structure that is common in educational and psychological data. These models provide information at both the individual (e.g., examinee or student) level and group (e.g., teacher or school) level. Student-level latent classes capture the association between the responses at the student-level unit while school-level latent classes capture the association between the students within school-level units (Cho & Cohen, 2010; Vermunt, 2003). If the hierarchical structure is ignored in the model, the model will not accurately detect latent classes. The MMixIRT model has been found to be useful in solving psychometric problems faced with standard DIF detection methods (Cho & Cohen, 2010) as well as several practical problems in testing applications. As an example, considering the Level 2 units as teachers, then teachers might be classified into latent classes which differed in their instructional emphasis on topics or in their use of particular sets of problem solving strategies (see Fagginger Auer, Hickendorff, van Putten, Béguin, & Heiser, 2016 for an example with multilevel latent class models).

MMixIRT models have been used in several studies including Bacci and Ginaldi (2015); Cho and Cohen (2010); Finch and Finch (2013); Jilke, Meuleman, and van de Walle (2015); Tay, Diener, Drasgow, and Vermunt (2011); Varriale and Vermunt (2012); and Vermunt (2008). Except for Cho and Cohen (2010), all of these studies used maximum likelihood estimation (MLE). These studies also varied with respect to the MixIRT models considered. Vermunt (2008) provided a maximum likelihood framework for estimating MMixIRT models and included an application to educational data with a multilevel mixture Rasch model (MMixRM) and a multilevel mixture two-parameter logistic IRT model. Varriale and Vermunt (2012) proposed a multilevel factor analysis for multilevel data sets with categorical latent variables at the higher level. Tay et al. (2011) introduced a multilevel mixed-measurement IRT model with an application to a data set on self-reported emotions. Bacci and Ginaldi (2015) classified university courses into homogeneous classes based on students' satisfactions using a two-level multidimensional mixture IRT model in which students represent the first-level units and courses are the second-level units. Bacci and Ginaldi used ordered polytomous items rather than dichotomous items. Similar to Cho and Cohen (2010), Finch and Finch (2013) presented an application of a multilevel multidimensional MixIRT model for DIF detection. Jilke et al. (2015) applied a MMixIRT model to test and correct for measurement nonequivalence within a single model. Jilke et al. compared the MMixIRT model using a multigroup confirmatory factor analysis model, a typical approach for determining measurement nonequivalence.

Exploratory applications of MixIRT models and their multilevel extensions focus on the detection of the best fitting model given the data. Unlike multiple group IRT models, the number of latent classes is not typically known a priori in an exploratory MixIRT model analysis. In such a study, multiple MixIRT models are fit to the data to determine the best fitting among the models being considered. The process of selecting the best fitting model is then typically done using some model selection criterion as well as a theoretical basis for each model being considered. Several model selection techniques have been proposed in the literature that can be broadly classified as either likelihood ratio based statistics or information criterion indices (Zucchini, 2000). The former is appropriate with nested models while the latter can be used with both nested and nonnested models. As is the case with MixIRT models, likelihood ratio based statistics are not appropriate for model selection in MMixRM (Nylund, Asparouhov, & Muthén, 2007). Information criterion indices were used in this study.

Information criterion indices have been used for selection with MixIRT models (Li, Cohen, Kim, & Cho, 2009) as well as MMixIRT models (Cho & Cohen, 2010) for comparing the relative fit of nonnested models based primarily on different numbers of classes. Results of simulation studies (e.g., Li et al., 2009; Preinerstorfer & Formann, 2011) have demonstrated that the Bayesian information criterion (BIC; Schwarz, 1978) performed well for model selection with dichotomous MixIRT models. Although consistent Akaike information criterion (CAIC; Bozdogan, 1987) and BIC have been used for model selection with MMixIRT models, the sample size to be used for these indices is not yet clear (Varriale & Vermunt, 2012). Lukociene and Vermunt (2010) and Lukociene, Varriale, and Vermunt (2010) suggested that the number of higher level units should be the preferred sample size for BIC and CAIC calculation for multilevel latent class models. Research on this issue, however, has not yet been reported in the literature for MMixIRT models.

In this study, the authors investigate the accuracy of four model selection indices commonly used in exploratory studies of MMixIRT models: Akaike information criterion (AIC; Akaike, 1974), CAIC, BIC, and sample-size adjusted Bayesian information criterion (SABIC; Sclove, 1987). Furthermore, the authors address the issue of which sample size should be used with MMixIRT models, as it is not clear whether number of students or number of schools should be used in the calculation of these three fit indices (i.e., BIC, CAIC, and SABIC) for the selection of multilevel models (Lukociene et al., 2010; Vermunt, 2011). A simulation study was first presented to explore the relative behavior of these indices under practical testing conditions. Next, an example was provided of the use of the indices on a set of real data to illustrate the behavior of the indices. In the next section, an overview of the MMixIRT model framework was provided, without loss of generality, in the context of the MMixRM.

MMixRM

The MMixRM (Asparouhov & Muthén, 2008; Cho & Cohen, 2010; Vermunt, 2007) has been proposed to deal with heterogeneity in multilevel data sets. As with the mixture Rasch model (MixRM), the MMixRM classifies respondents into different latent classes at multiple levels. In a school context, for example, since students can be viewed as nested within schools, there could be both student-level (i.e., Level 1) and school-level (Level 2) latent classes. The conditional probability of a correct response in the MMixRM can be given as

$$P(x_{ijt} = 1 | g, k, \theta_{jtgk}) = \frac{1}{1 + \exp[-(\theta_{jtgk} - \beta_{igk})]}, \quad (1)$$

where $g = 1, \dots, G$ is an index for student-level latent classes, $k = 1, \dots, K$ is an index for school-level latent classes, $j = 1, \dots, J$ is an index for examinees, $t = 1, \dots, T$ is an index for schools, for a test of $i = 1, \dots, I$ items, β_{igk} is the difficulty of item i for latent classes g and k , and θ_{jtgk} is the ability of examinee j in school t and in latent classes g and k . Abilities θ_{jtgk} have mean μ_{gk} and variance σ_{gk}^2 with normal distribution (i.e., $\theta_{jtgk} \sim N(\mu_{gk}, \sigma_{gk}^2)$).

As shown in Cho (2007), θ_{jtgk} is reparameterized into $\sqrt{\sigma_{gk}^2} \cdot \eta_{jtgk}$ to anchor the metric. The η_{jtgk} is set to $\eta_{jtgk} \sim \text{Normal}(\mu_{gk}, 1)$, where $\eta_{jtgk} = \theta_{jtgk} / \sqrt{\sigma_{gk}^2}$. For metric anchoring, the mean of μ_{gk} is set to 0 for the reference group (in this study, $\mu_{11} = 0$). The means for the remaining latent classes are set to μ_{gk} , that is, they are to be estimated. Like the MixRM, each latent class has a different set of item difficulty and ability parameter estimates. In addition, mixture proportions are estimated for both student-level ($\pi_{g|k}$) and school-level (π_k) classes.

Estimation of Model Parameters

Bayesian estimation of the MMixRM as described in Cho and Cohen (2010) requires combination prior knowledge with the data to obtain the posterior distribution. The following priors for Bayesian estimation of a MMixRM were used as described in Cho and Cohen (2010):

$$\begin{aligned} k &\sim \text{Categorical}(\{1, \dots, K\}, (\pi_1, \dots, \pi_K)); \\ g|k &\sim \text{Categorical}(\{1, \dots, G\}, (\pi_{1|k}, \dots, \pi_{G|k})); \\ \mu_{gk}|g, k &\sim \text{Normal}(0, 1) \text{ if } (g, k) \neq (1, 1) \text{ and } \mu_{11} = 1; \\ \sigma_{gk}|g, k &\sim \text{Normal}(0, 1)I(0,); \\ \beta_{igk}|g, k &\sim \text{Normal}(0, 1), i = 1, \dots, I, \end{aligned}$$

where $I(0,)$ indicates that observations of σ were sampled above 0. The probabilities of mixtures were modeled using a Dirichlet distribution as the conjugate prior of the parameters of the multinomial distribution, where $\sum_{g=1}^G \pi_{g|k} = 1$ for all school-level latent class ks with the proportion of π_k and G indicates the number of student-level classes. The Dirichlet distribution with Gamma sampling was used as a prior of π_k as the Gamma distribution can be used to sample G independent random variables from Dirichlet distribution (Gelman, Carlin, Stern, & Rubin, 2003).

Model Selection

Model selection for nonnested models or for models that do not follow regularity conditions can be aided using information criterion indices. The performance of four information criterion indices (described below) were investigated in this study. Information criterion indices are based on some form of penalization of the loglikelihood. The penalization is used to adjust for the selection of overparameterized models. Each information index applies a different penalty function. Let L be the likelihood function obtained from MLE and P be the penalty term. The following is a general form for these indices:

$$-2 \log L + P. \quad (2)$$

As each information index applies a different penalty function to the same loglikelihood, it is possible to obtain solutions for the different indices using the same data. Following the general form in Equation (2), AIC can be calculated as

$$\text{AIC} = -2 \log L + 2d, \quad (3)$$

where d is the number of estimated parameters and $2d$ is the penalty function.

The lack of a sample size penalty function leads to some inconsistency in the performance of AIC as it tends to overestimate the correct number of classes (McLachlan & Peel, 2000). Furthermore, AIC has a tendency to select more complex mixture IRT models compared with other information indices (Li et al., 2009).

As AIC is not consistent (Sclove, 1987; Tofighi & Enders, 2008), Bozdogan (1987) proposed CAIC as an alternative to AIC. The penalty function in CAIC includes both the number of parameters and the sample size:

$$\text{CAIC} = -2 \log L + d[\ln(N) + 1]. \quad (4)$$

The penalty function in BIC also includes the number of parameters and the log of the sample size. BIC can be described as

$$\text{BIC} = -2 \log L + d \cdot \ln(N), \quad (5)$$

where d is the number of free parameters estimated and $\ln(N)$ is the natural log of the total sample size, N . Inclusion of sample size in the BIC formula makes BIC behave different than AIC. As the sample size increases BIC consistently selects the true model, when it is under consideration. Unlike BIC, AIC tends to select more complex model as sample size increases, even when the true model is in the candidate model set. As a result, BIC tends to select less complex models than AIC (Congdon, 2003; Sclove, 1987). If the true model is not under consideration, then AIC is asymptotically optimal in a certain sense. As Shao (1997) notes, the usefulness of AIC and BIC may depend on the structure of true model (i.e., simple vs. complex). In the case of simple model structure, BIC is expected to perform well; AIC is expected to work better, when the true model has a complex structure. A number of factors including the loss function, study design, and research question have an affect on the performance of AIC and BIC indices (Vrieze, 2012). In addition, class separation and class proportions have been shown to affect the usefulness of these two fit indices within latent variable models context (Lubke & Neale, 2006; Nylund et al., 2007; Vrieze, 2012). BIC has been found to be more effective than AIC for selection of dichotomous MixIRT models (Li et al., 2009; Preinerstorfer & Formann, 2011; Sen, 2014; Sen, Cohen, & Kim, 2016).

Sclove (1987) proposed SABIC, a sample size adjusted version of BIC, by replacing the sample size (N) with $[N^* = (N + 2)/24]$, thereby reducing the penalty due to sample size. BIC and SABIC indices are also based on the maximum likelihood, instead of on an estimate based on a Bayesian posterior. The Bayesian term in the name BIC comes from assigning a prior to the models in the model space (Schwarz, 1978; Sclove, 1987). SABIC has been shown to work better with large number of parameters or small sample sizes (Yang, 2006). Previous research has shown that SABIC may be useful for model selection in latent variable mixture models (Henson, Reise, & Kim, 2007; Tofighi & Enders, 2008; Yang, 2006).

The number of students (i.e., total sample size) is typically used for single level models. There is some question, however, whether this is appropriate for multilevel models. In the context of multilevel latent class models, Lukociene et al. (2010) and Vermunt (2011) suggested using the number of groups at Level 2 as the “sample size.” In a two-level model, this would mean that the appropriate “sample size” for the penalty function would be the number of groups at Level 2. In the present study, these two definitions of sample size were investigated. AIC, BIC, CAIC, and SABIC were used to indicate the information indices with total sample size (the number of students) in the penalty function and BIC2, CAIC2, and SABIC2 to indicate the indices with level two sample size (the number of schools) in the penalty function.

Bayesian estimation produces a different likelihood than that obtained with MLE. The likelihood function is obtained from the posterior distribution. As described in Cho and Cohen (2010), the likelihood function of the MMixRM is as follows:

$$L(g, k, \theta_{jtgk}) = \prod_{i=1}^I \prod_{j=1}^J \left[\left\{ \sum_{k=1}^K \sum_{g=1}^G \pi_k \cdot \pi_{g|k} P(x_{ijt} = 1 | g, k, \theta_{jtgk}) \right\}^{u_{ij}} \cdot \left\{ 1 - \sum_{k=1}^K \sum_{g=1}^G \pi_k \cdot \pi_{g|k} P(x_{ijt} = 1 | g, k, \theta_{jtgk}) \right\}^{1-u_{ij}} \right]^{L_{jtgk}}, \quad (6)$$

where u_{ij} are dichotomously scored responses as 0 and 1, ζ_{jgk}^l is 1 if the examinee j is from mixtures g and k and $\zeta_{jgk}^l = 0$ otherwise at an iteration l .

Information indices can be calculated in Bayesian analyses using either averages of deviances, D , over a Markov chain Monte Carlo (MCMC) chain or the deviance at the mean of the posterior (Congdon, 2003). Averaging the deviance can be done attaching $(D^{(l)})$ to parameters $\theta^{(l)}$ at each iteration. In this study, Bayesian estimation replaced the MLE-based deviance (i.e., $-2 \log L$) with the posterior mean of the deviance $\overline{D(\xi)}$, where ξ represents all of the estimated parameters including item parameter estimates, mixture proportions and latent class means. Thus, the definitions of relative fit indices in this study are as follows:

$$AIC = -2D^t + 2d, \quad (7)$$

$$BIC = -2D^t + d \ln(N), \quad (8)$$

$$CAIC = -2D^t + d[\ln(N) + 1], \quad (9)$$

$$SABIC = -2D^t + d[\ln((N+2)/24)]. \quad (10)$$

Simulation Study

A simulation study was conducted to investigate the performance of the four different information criterion indices (AIC, BIC, CAIC, and SABIC) for use in selection of the best MMixRM. BIC, CAIC, and SABIC were calculated with two different types of sample sizes for the penalty function.

Design of the Simulation Study

The following factors were manipulated in the simulation study: size of sample of students at Level 1 (1,500 and 3,000 corresponding), size of sample at Level 2 (50 and 100 schools in which each schools has 30 students), test length (10 and 20 items), number of student-level latent classes (G = one, two, and three student-level classes) and number of school-level latent classes (K = one, two, and three school-level classes). Four combinations of student- and school-level classes were simulated: one student level and one school level (G1K1), two student levels and two school levels (G2K2), three student levels and two school levels (G3K2), and three student levels and three school levels (G3K3). The G1K1 condition is the testing condition usually assumed for an IRT model. In this study, it can be considered a type of baseline. The generating parameters for the study were obtained from estimates from empirical data collected as part of a larger project (Raczynski, Alagoz, Thompson, & Pollack, 2015).

A Dirichlet distribution with Gamma sampling on the probabilities of mixtures was used for data generation. The latent classes were generated at both the student level and the school level using equal mixing proportions. For example, both school-level latent classes were simulated to have mixing proportions of .5 for G2K2 conditions. All data sets were generated using WinBUGS 1.4 (Spiegelhalter, Thomas, & Best, 2003) software as described in Cho and Cohen (2010).

Estimation

A MCMC algorithm was used for estimation of the MMixRM as implemented in the WinBUGS 1.4 software. Starting values for all parameters were randomly generated in the WinBUGS software. The prior distributions presented above were used. As described in Cho (2007), the

identification problem of the ability distribution (Choi, 2014) was solved using metric identification (anchoring the metric with respect to the ability distribution).

An adaptive rejection sampling algorithm was used with 10,000 iterations for the burn-in and 10,000 iterations for the post burn-in iterations for the MCMC analyses. Several criteria have been used to examine convergence in MCMC estimation (Cowles & Carlin, 1996; Gelman & Rubin, 1992). The R package “convergence diagnosis and output analysis for MCMC” (CODA; Plummer, Best, Cowles, & Vines, 2006) was used in this study. In addition, density plots and the ratio of the Monte Carlo standard error to the standard deviation of the parameter were also considered. For complex models with many parameters, it can be impractical to check convergence for every parameter, so monitoring of a most relevant parameter may be a more practical alternative. Loglikelihood values were monitored in this study to help with determining the number of burn-in and post burn-in iterations. Only data sets with the large sample size in the simulation study (i.e., 3,000 simulated examinees) and the long test length (i.e., 20 items), therefore, were used for convergence assessment. The number of burn-in and post burn-in iterations were determined by examining the Gelman–Rubin convergence statistics. As calculation of Gelman–Rubin convergence statistics requires at least two chains, MCMC analyses for each condition in this study were done with two parallel chains.

A sample analysis is presented for the loglikelihood value to provide a brief example of what was done to determine convergence. For this sample analysis, 15,000 iterations were used for both burn-in and post burn-in to determine the appropriate number of burn-in iterations. Figure 1 presents a Gelman–Rubin shrinkage factor plot, a history plot, a density plot, and an autocorrelation plot for the loglikelihood value for the data set generated with the G3K3 MMixIRT model. The convergence of an MCMC estimation can be assessed using these plots. The Gelman–Rubin shrinkage plot is used to show that multiple chains converges to the same target stationary distribution. The upper plot in Figure 1 shows that both the median and 97.5% quantile for the loglikelihood stabilize around a value of 1.0 after about the first 18,000 iterations. As this diagnostic is calculated from the second half of each chain, it suggests that convergence was achieved after approximately 9,000 iterations. The history plot helps to understand if the two chains overlap and if the chains are mixing well. As can be seen in the second plot in Figure 1, the chains appear to mix well with each other despite starting at different points. The third plot in Figure 1 is the density plot for the loglikelihood estimate. This is simply a histogram of the distributions of every draw. Converged chains produce a uni-modal (e.g., bell-shaped) density plot as shown in Figure 1. The autocorrelation plot in Figure 1 shows the correlation between each draw and previous draws. This plot shows a decay which indicates good convergence. Finally, the loglikelihood was estimated to be $-18,010.0$, and the Monte Carlo error of 4.896 was less than 5% of the sample standard deviation (Spiegelhalter et al., 2003). From this, the authors conclude that the MCMC chain converged. Based on these preliminary results, a conservative number of 10,000 burn-in iterations and 10,000 post burn-in iterations were selected for the MixRM conditions.

A total of 50 replications were simulated for each of the 2 Total Sample Sizes $\times$ 2 Test Lengths $\times$ 4 Combinations of student- and school-level latent classes = 16 conditions. Each condition was estimated with six different models: G1K1, G2K2, G3K2, G3K3, G4K2 (four student levels and two school levels), and G4K3 (four student levels and three school levels). A correct detection rate was calculated for each of the information criterion indices for each condition. The correct detection rate was defined as the correct detection of the simulated model with the correct number of student- and school-level latent classes.

Figure 1. Convergence plots for the sample loglikelihood parameter.

Simulation Study Results

Recovery of item difficulty parameters was assessed using root mean square error (RMSE) and bias between the generating parameters and the parameter estimates of the true-generated model. Bias and RMSE values were calculated using the following formulas:

$$\text{RMSE}_{(\beta_{igk})} = \sqrt{\frac{\sum_{i=1}^I \sum_{g=1}^G \sum_{k=1}^K (\beta_{igk} - \hat{\beta}_{igk})^2}{IGK}}, \quad (11)$$

$$\text{Bias}_{(\beta_{igk})} = \frac{\sum_{i=1}^I \sum_{g=1}^G \sum_{k=1}^K (\beta_{igk} - \hat{\beta}_{igk})}{IGK}, \quad (12)$$

where I denotes number of iterations, G indicates the number of student level classes and K is the index for the number of school level classes. Here β_{igk} is the generating item difficulty parameter values and $\hat{\beta}_{igk}$ is the item difficulty parameter which is estimated as the mean of the posterior distribution from the generated model.

Table 1 presents mean RMSE and bias values for each condition. Mean RMSE values for item difficulty estimates ranged from 0.273 to 0.766. Mean bias values ranged between -0.584 and 0.575 . The G1K1 data generation conditions yielded the lowest mean RMSE and mean bias values. G3K2 conditions had the highest mean RMSE values (ranging from 0.521 to 0.766) and mean bias values (ranging from 0.375 to 0.575).

Table 1. Recovery Statistics for Generated Data Conditions.

Generating Model	Test Length	Sample Size	RMSE	Bias
G1K1	10	1,500	0.273	-0.155
G1K1	10	3,000	0.341	-0.148
G1K1	20	1,500	0.312	-0.143
G1K1	20	3,000	0.388	-0.119
G2K2	10	1,500	0.532	-0.418
G2K2	10	3,000	0.710	-0.584
G2K2	20	1,500	0.658	-0.402
G2K2	20	3,000	0.766	-0.561
G3K2	10	1,500	0.521	-0.375
G3K2	10	3,000	0.570	-0.416
G3K2	20	1,500	0.756	0.510
G3K2	20	3,000	0.670	0.575
G3K3	10	1,500	0.564	-0.517
G3K3	10	3,000	0.531	-0.497
G3K3	20	1,500	0.397	0.196
G3K3	20	3,000	0.375	0.136

Note. RMSE = root mean square error; G1K1 = one student level and one school level; G2K2 = two student levels and two school levels; G3K2 = three student levels and two school levels; G3K3 = three student levels and three school levels.

The frequencies of correct model selections are shown in Tables 2, 3, 4, and 5 for each of the information indices using the total and Level 2 sample sizes. The first and second columns in these tables present the number of items and number of examinees for each condition. The numbers of correct detections for each information index are indicated in bold. A two proportion z test for each condition was used to test whether any of the fit indices performed significantly better. Proportions of correct detections obtained from each fit index were compared. A two-tailed z test and α level of .05 were used to test the null hypothesis that there was no difference between the two proportions from the different fit indices. Based on this z test criterion, significant differences were found in the performances of the two fit indices, when there were 10 or more correct detection differences.

Results in Tables 2, 3, 4, and 5 show that AIC performed poorly for most conditions, except for the G1K1 condition with 20 items and 1,500 examinees. Correct detection frequencies ranged between 0 and 12 out of 50 replications in the remaining conditions. Although AIC performed better in the 50 School $\times$ 1,500 Examinees Conditions than 100 School $\times$ 3,000 Examinees Conditions, AIC most often chose other models than the true data-generating model. These tended to be more complex models than the true generating models.

Correct Detections Using Total Sample Size

Results for correct detections for total sample size for G1K1 were good for both BIC and CAIC and relatively good for SABIC. These results were relatively better than results based on number of schools. In the G1K1 conditions, the true model was chosen correctly 49 to 50 times by BIC; the number of correct detections was 50 for CAIC index for all item and sample size conditions. The true model was chosen correctly 39 to 50 times by SABIC, indicating lower rates than both BIC and CAIC. As can be seen in Table 2, the SABIC index incorrectly selected G2K2 as the true model in a number of the G1K1 conditions.

Table 2. Detection Frequencies for G1K1 Data Generation Conditions.

Test length	Total sample size	Model	Total sample size penalty function				Level 2 sample size penalty function		
			AIC	CAIC	BIC	SABIC	CAIC2	BIC2	SABIC2
10	1,500	G1K1	6	50	49	39	45	34	0
10	1,500	G2K2	1	0	1	6	3	7	0
10	1,500	G3K2	6	0	0	3	1	6	1
10	1,500	G3K3	3	0	0	0	0	0	1
10	1,500	G4K2	23	0	0	2	1	3	12
10	1,500	G4K3	11	0	0	0	0	0	36
10	3,000	G1K1	10	50	50	49	50	47	5
10	3,000	G2K2	10	0	0	1	0	3	9
10	3,000	G3K2	10	0	0	0	0	0	4
10	3,000	G3K3	6	0	0	0	0	0	21
10	3,000	G4K2	12	0	0	0	0	0	11
10	3,000	G4K3	2	0	0	0	0	0	0
20	1,500	G1K1	35	50	50	50	50	50	1
20	1,500	G2K2	8	0	0	0	0	0	6
20	1,500	G3K2	6	0	0	0	0	0	6
20	1,500	G3K3	0	0	0	0	0	0	4
20	1,500	G4K2	1	0	0	0	0	0	13
20	1,500	G4K3	0	0	0	0	0	0	20
20	3,000	G1K1	12	50	50	48	50	48	6
20	3,000	G2K2	9	0	0	2	0	2	8
20	3,000	G3K2	11	0	0	0	0	0	4
20	3,000	G3K3	2	0	0	0	0	0	21
20	3,000	G4K2	14	0	0	0	0	0	11
20	3,000	G4K3	2	0	0	0	0	0	0

Note. G1K1 = one student level and one school level; AIC = Akaike information criterion; CAIC = consistent AIC; BIC = Bayesian information criterion; SABIC sample-size adjusted BIC; G2K2 = two student levels and two school levels; G3K2 = three student levels and two school levels; G3K3 = three student levels and three school levels; G4K2 = four student levels and two levels; G4K3 = four student levels and three school levels.

Detection frequencies for G2K2 conditions are presented in Table 3. As can be seen in Table 2, use of the total sample size appeared to be relatively effective for the G2K2 conditions. Large frequencies of correct detections for the BIC index for the G2K2 condition were observed. G2K2 10-item conditions were correctly recovered 35 times for 1,500 examinees; for the 20-item condition, the number of correct detections was greater than 41. In the G2K2 10-item conditions, however, recovery was less successful with 23 out of 50 correct detections for conditions with 3,000 examinees. The frequencies of correct detections for CAIC were also large for most of the G2K2 conditions. Furthermore, BIC and CAIC incorrectly selected the G4K2 and G3K2 as the best fitting models in the 10-item conditions. In the G1K1 condition, the estimated model was incorrectly selected by BIC and CAIC for the 20-item conditions. Performance of the SABIC information criterion was slightly lower for G2K2 conditions than in the G1K1 conditions. SABIC selected the true generating model more often in the G2K2 $\times$ 20-Item Conditions than in the G2K2 $\times$ 10-Item Conditions. However, SABIC incorrectly selected the correct model less frequently in the G4K2 and G3K2 conditions.

Detection frequencies for G3K2 conditions are presented in Table 4 with correct detection frequencies based on BIC, CAIC, and SABIC indices ranging from 5 to 29. BIC and CAIC indices had the highest number of correct detection frequencies (i.e., 29) for the 10 Items $\times$

Table 3. Detection Frequencies for G2K2 Data Generation Conditions.

Test length	Total sample size	Model	Total sample size penalty function				Level 2 sample size penalty function		
			AIC	CAIC	BIC	SABIC	CAIC2	BIC2	SABIC2
10	1,500	G1K1	0	3	2	0	0	0	0
10	1,500	G2K2	0	42	35	13	21	10	0
10	1,500	G3K2	11	5	11	25	22	28	1
10	1,500	G3K3	2	0	0	1	0	1	2
10	1,500	G4K2	18	0	2	11	7	11	10
10	1,500	G4K3	19	0	0	0	0	0	37
10	3,000	G1K1	0	3	3	0	0	0	0
10	3,000	G2K2	1	28	23	6	12	6	1
10	3,000	G3K2	2	15	17	15	17	13	1
10	3,000	G3K3	2	0	0	0	0	0	2
10	3,000	G4K2	29	4	7	28	21	29	26
10	3,000	G4K3	16	0	0	1	0	2	20
20	1,500	G1K1	0	15	9	1	3	1	0
20	1,500	G2K2	12	35	41	41	42	39	1
20	1,500	G3K2	9	0	0	5	3	5	2
20	1,500	G3K3	2	0	0	0	2	0	3
20	1,500	G4K2	24	0	0	3	0	5	16
20	1,500	G4K3	3	0	0	0	0	0	28
20	3,000	G1K1	0	9	9	3	3	3	0
20	3,000	G2K2	1	40	39	25	34	23	1
20	3,000	G3K2	11	1	2	15	12	17	5
20	3,000	G3K3	2	0	0	0	0	0	2
20	3,000	G4K2	29	0	0	7	1	7	30
20	3,000	G4K3	7	0	0	0	0	0	12

Note. G2K2 = two student levels and two school levels; AIC = Akaike information criterion; CAIC = consistent AIC; BIC = Bayesian information criterion; SABIC sample-size adjusted BIC; G1K1 = one student level and one school level; G3K2 = three student levels and two school levels; G3K3 = three student levels and three school levels; G4K2 = four student levels and two school levels; G4K3 = four student levels and three school levels.

3,000 Examinees Condition. Correct detection rates for BIC, CAIC, and SABIC tended to decrease as model complexity increased. AIC index selected more complex models most of the times. Overall, CAIC and BIC performed better for G3K2 than the other indices. However, all fit indices incorrectly selected the G4K2 and G2K2 (the closest models to true model) as the best fitting models under most of the G3K2 conditions.

Detection frequencies for G3K3 conditions are presented in Table 5. None of the information indices selected the generating model for any of the G3K3 conditions. Correct detection frequencies ranged from 0 to 8 for G3K3 conditions. The AIC index had the highest correct detection frequencies (5 and 8 out of 50) for G3K3 conditions with 20 items. As can be seen in Table 5, BIC and CAIC incorrectly selected simpler models (G2K2 and G3K2) whereas AIC incorrectly selected more complex models (i.e., G4K2 and G4K3) for most of the G3K3 conditions. SABIC tended to incorrectly select the nearest models (i.e., G3K2 and G4K2) to the true generating model (i.e., G3K3).

CAIC performed much better than AIC, BIC, and SABIC (see Table 2), whereas BIC performed better than AIC and SABIC under G1K1 conditions with 10 Items $\times$ 1,500 Examinees. CAIC, BIC, and SABIC performed better than AIC index under the remaining G1K1 conditions. As can be seen in Table 3, CAIC and BIC performed better than AIC and SABIC whereas

Table 4. Detection Frequencies for G3K2 Data Generation Conditions.

Test length	Total sample size	Model	Total sample size penalty function				Level 2 sample size penalty function		
			AIC	CAIC	BIC	SABIC	CAIC2	BIC2	SABIC2
10	1,500	G1K1	0	0	0	0	0	0	0
10	1,500	G2K2	2	29	25	6	13	5	2
10	1,500	G3K2	12	18	20	21	22	21	5
10	1,500	G3K3	1	0	0	2	0	1	1
10	1,500	G4K2	16	3	5	18	14	19	11
10	1,500	G4K3	19	0	0	3	1	4	31
10	3,000	G1K1	0	0	0	0	0	0	0
10	3,000	G2K2	1	7	4	2	2	2	1
10	3,000	G3K2	8	29	29	16	18	16	6
10	3,000	G3K3	3	1	1	1	2	1	3
10	3,000	G4K2	21	13	16	28	26	27	19
10	3,000	G4K3	17	0	0	3	2	4	21
20	1,500	G1K1	0	9	3	0	0	0	0
20	1,500	G2K2	3	36	39	21	29	20	0
20	1,500	G3K2	4	5	8	14	13	13	1
20	1,500	G3K3	4	0	0	3	1	3	4
20	1,500	G4K2	29	0	0	12	7	14	18
20	1,500	G4K3	10	0	0	0	0	0	27
20	3,000	G1K1	0	0	0	0	0	0	0
20	3,000	G2K2	1	32	29	16	19	13	1
20	3,000	G3K2	3	15	15	8	12	8	3
20	3,000	G3K3	9	0	1	4	3	6	7
20	3,000	G4K2	16	3	5	22	16	22	13
20	3,000	G4K3	21	0	0	0	0	1	26

Note. G3K2 = three student levels and two school levels; AIC = Akaike information criterion; CAIC = consistent AIC; BIC = Bayesian information criterion; SABIC sample-size adjusted BIC; G1K1 = one student level and one school level; G2K2 = two student levels and two school level; G3K3 = three student levels and three school levels; G4K2 = four student levels and two school levels; G4K3 = four student levels and three school levels.

SABIC performed better than AIC under G2K2 conditions. CAIC and BIC performed better than AIC and SABIC (see Table 4). SABIC performed better than AIC under G3K2 conditions with 10 Items $\times$ 3,000 examinees. In addition, SABIC performed better than AIC under G3K2 conditions with 20 Items $\times$ 1,500 Examinees. However, there were no substantial differences among the performances of four total sample size-based fit indices under the remaining G3K2 conditions. Likewise, as can be seen in Table 5, there were no substantial differences among the performances of the seven fit indices under any of G3K3 conditions.

Correct Detections Using Sample Size at Level 2

The last three columns of Tables 2, 3, 4, and 5 present correct detection frequencies for BIC2, CAIC2, and SABIC2. Both CAIC2 and BIC2 performed relatively well for the simpler G1K1 conditions (i.e., for the conditions with one class at Level 2). The correct detection frequencies for CAIC2 and BIC2 ranged from 34 to 50. SABIC2 had the poorest performance under G1K1 conditions, incorrectly selecting the most complex models (G3K3, G4K2, and G4K3). SABIC2 incorrectly selected G3K3 and G4K2 models in the 3,000-examinee conditions whereas the most incorrectly detected models were in the 1,500-examinee conditions (see Table 2).

Table 5. Detection Frequencies for G3K3 Data Generation Conditions.

Test length	Total sample size	Model	Total sample size penalty function				Level 2 sample size penalty function		
			AIC	CAIC	BIC	SABIC	CAIC2	BIC2	SABIC2
10	1,500	G1K1	0	4	3	0	0	0	0
10	1,500	G2K2	0	32	23	3	9	1	0
10	1,500	G3K2	7	13	23	29	31	28	1
10	1,500	G3K3	3	0	0	3	1	3	2
10	1,500	G4K2	17	1	1	13	9	16	5
10	1,500	G4K3	23	0	0	2	0	2	42
10	3,000	G1K1	0	0	0	0	0	0	0
10	3,000	G2K2	0	12	6	0	0	0	0
10	3,000	G3K2	2	28	25	12	19	9	0
10	3,000	G3K3	0	0	0	0	0	0	1
10	3,000	G4K2	21	10	19	32	29	35	16
10	3,000	G4K3	27	0	0	6	2	6	33
20	1,500	G1K1	0	2	1	0	0	0	0
20	1,500	G2K2	0	40	39	15	21	14	2
20	1,500	G3K2	12	7	9	21	21	21	4
20	1,500	G3K3	5	0	0	3	2	3	0
20	1,500	G4K2	13	1	1	10	6	11	8
20	1,500	G4K3	20	0	0	1	0	1	36
20	3,000	G1K1	0	0	0	0	0	0	0
20	3,000	G2K2	0	32	26	13	15	11	0
20	3,000	G3K2	4	15	15	13	13	13	2
20	3,000	G3K3	8	0	0	2	1	2	7
20	3,000	G4K2	19	3	9	21	21	20	14
20	3,000	G4K3	19	0	0	1	0	4	27

Note. G3K3 = three student levels and three school levels; AIC = Akaike information criterion; CAIC = consistent AIC; BIC = Bayesian information criterion; SABIC sample-size adjusted BIC; G1K1 = one student level and one school level; G2K2 = two student levels and two school levels; G3K2 = three student levels and two school levels; G4K2 = four student levels and two school levels; G4K3 = four student levels and three school levels.

Similarly, CAIC2 and BIC2 indices performed well for the G2K2 20-item test conditions. The performance of SABIC2, however, was poor with correct detections ranging from 0 to 7. This was markedly different than the performance of SABIC (i.e., using total sample size). As in the G1K1 conditions, SABIC2 incorrectly selected the most complex models (G3K3, G4K2, and G4K3) under G1K1 conditions (see Table 3).

As shown in Table 4, the results for Level-2 sample sizes under G3K2 data conditions were similar to those under G1K1 and G2K2 conditions. CAIC2 and BIC2 indices performed well and SABIC2 performed poorly in selecting the true generating model (i.e., G3K2). Similarly, SABIC2 incorrectly selected the more complex models (G3K3, G4K2, and G4K3) under G3K2 conditions (see Table 3).

None of the indices using sample size at Level 2 performed well for any of the G3K3 conditions. As can be seen in Table 6, Table 5, BIC2 and CAIC2 incorrectly selected G2K2, G3K2, and G4K2 models whereas SABIC2 incorrectly selected more complex models (i.e., G4K2 and G4K3) for most of the G3K3 conditions.

As can be seen in Table 2, 3, and 4, CAIC2 and BIC2 performed much better than SABIC2 under G1K1, G2K2, and G3K2 conditions, respectively. As shown in Table 5, there were no

Table 6. Model Fit Statistics for Empirical Data.

Model	-2loglikelihood	AIC	BIC	CAIC	SABIC	BIC2	CAIC2	SABIC2
G1K1	53,917.31	53,945.31	54,034.11	54,048.11	53,989.62	53,969.63	53,983.63	53,925.79
G2K2	52,359.47	52,475.47	52,843.35	52,901.35	52,659.06	52,576.25	52,634.25	52,394.63
G3K2	51,407.88	51,583.88	52,142.05	52,230.05	51,862.43	51,736.80	51,824.80	51,461.22
G3K3	51,124.34	51,390.34	52,233.94	52,366.94	51,811.32	51,621.45	51,754.45	51,204.96
G4K2	51,003.11	51,239.11	51,987.57	52,105.57	51,612.61	51,444.16	51,562.16	51,074.63
G4K3	50,537.32	50,893.32	52,022.35	52,200.35	51,456.74	51,202.63	51,380.63	50,645.22
G4K4	50,580.00	51,056.00	52,565.60	52,803.60	51,809.33	51,469.57	51,707.57	50,724.26
G5K2	50,600.00	50,896.00	51,834.74	51,982.74	51,364.46	51,153.18	51,301.18	50,689.71
G5K3	50,520.00	50,966.00	52,380.45	52,603.45	51,671.85	51,353.50	51,576.50	50,655.17
G5K4	50,280.00	50,876.00	52,766.17	53,064.17	51,819.25	51,393.83	51,691.83	50,460.63
G5K5	50,100.00	50,846.00	53,211.88	53,584.88	52,026.64	51,494.15	51,867.15	50,326.09

Note. AIC = Akaike information criterion; CAIC = consistent AIC; BIC = Bayesian information criterion; SABIC sample-size adjusted BIC; G1K1 = one student level and one school level; G2K2 = two student levels and two school levels; G3K2 = three student levels and two school levels; G3K3 = three student levels and three school levels; G4K2 = four student levels and two school levels; G4K3 = four student levels and three school levels. G4K4 = four student levels and four school levels; G5K2 = five student levels and two school levels; G5K3 = five student levels and three school levels; G5K4 = five student levels and four school levels; G5K5 = five student levels and five school levels.

major differences among the performances of three Level-2 fit indices under any of G3K3 conditions.

Empirical Illustration: Trends in International Mathematics and Science Study (TIMSS) 2011 Math and Science Assessment

The following real data example is used to illustrate model selection in the context of a MMixRM. Data for this example were taken from the mathematics section of the 2011 TIMSS (Foy, Arora, & Stanco, 2013). There were six booklets that were made publicly available (Booklets 1, 2, 5, and 6). For this example, responses of 4,200 eighth-grade students in 42 schools in the United States to Booklet 1 were sampled from the total number of U.S. eighth-grade students (10,477) administered the TIMSS 2011 mathematics exam. Booklet 1 was composed of 26 items including 12 multiple choice and 14 constructed response items. Only 12 multiple choice questions in Booklet 1 were analyzed as a dichotomous MMixIRT model was used in this study. The intraclass correlation (ICC, Raudenbush & Bryk, 2002) was used to determine the hierarchical structure of the data. The ICC for the multilevel IRT model was .139; that is, 13.9% of the total variation in ability estimates was explained at the school level, indicating a multilevel model was needed.

The following 11 multilevel models were estimated: G1K1, G2K2, G3K2, G3K3, G4K2, G4K3, G4K4 (four student levels and four school levels), G5K2 (five student levels and two school levels), G5K3 (five student levels and three school levels), G5K4 (five student levels and four school levels), and G5K5 (five student levels and five school levels). AIC, BIC, CAIC, and SABIC were calculated for each of these models. In addition to these indices, BIC2, CAIC2, and SABIC2 were calculated using Level 2 sample size. The smallest value of each information criterion index was taken as indicating the best fitting model.

Results of the Empirical Data

The MMixIRT models used in this empirical example were different than those used for the simulation study to determine the best model among a wide set of appropriate alternative models. The same estimation technique and priors applied in simulation study were also used in the empirical analyses. Several models were estimated using the same data, as the best model was not known a priori. G1K1, G2K2, G3K2, G3K3, G4K2, G4K3, G4K4, G5K2, G5K3, G5K4, and G5K5 were the estimated models with different number of classes at both student and school level. Model fit statistics for these 11 models are presented in Table 6. As shown in Table 6, all of the information criterion indices gave consistent results except for AIC and SABIC2 in terms of best fitting model. The G5K2 model was found to be best model with the smallest fit index values based on BIC, CAIC, SABIC, BIC2, and CAIC2 indices. However, AIC and SABIC2 indicated that G5K5 was the best model among alternatives. In view of these results, the authors can conclude that heterogeneity behind this real data can best be explained by the G5K2 model among the alternatives used for this example.

Discussion

Selection of the correct model among alternatives is important in selection of the best representation of the data among a set of candidate models. In the present study, information criterion indices were used to select the best model from a set of competing nonnested models. Results of the simulation study using total sample size indicated that CAIC and BIC indices performed

better than the SABIC and AIC under both 10-item and 20-item test lengths for G2K2 and for the 20-item test for G3K2 conditions. SABIC performed poorly, although still better than AIC. Consistent with previous research, AIC tended to select the more complex model regardless of the generating model compared with BIC, CAIC, and SABIC. None of these indices performed well for the most complex G3K3 model conditions. Results using Level 2 sample sizes suggested that BIC2 and CAIC2 were most effective at selecting the correct K1 models and to a lesser extent, the K2 models. None of the information criteria, however, were effective for the K3 models.

Different sample sizes for the penalty function did appear to have an effect on correct selections for the least complex multilevel model, that is, for G2K2. Results showed that the total sample size appeared to be the more appropriate sample size for the penalty function. This is somewhat different than the results reported by Lukociene and Vermunt (2010) and Lukociene et al. (2010). Lukociene and Vermunt (2010) and Lukociene et al. (2010) estimated multilevel latent class models. These models are less complex than multilevel IRT Models. Total sample size-based indices were found to be better than Level 2 sample size indices, but none of the total sample size-based indices performed well for the G3K3 conditions in this study. When the number of schools was used as the sample size and when it did not increase, BIC was not consistent as the number of individuals increased. It appears that using the number of individuals as the sample size yields better results.

The results of this study were consistent with previous research by Li et al. (2009) in that BIC was effective at selection of the correct model. Furthermore, in this study, CAIC was also as effective as BIC. Both indices, however, were only effective for the less complex models. This result suggests that CAIC and BIC performed well with simple models only because the penalty for model complexity is so high. In the case of the G1K1 condition, any index with a high penalty function would select the correct model as there was no model with less complexity. Success rates for selecting the correct model also varied with sample size.

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) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Sedat Sen <https://orcid.org/0000-0001-6962-4960>

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.
- Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27-51). Charlotte, NC: Information Age Publishing.
- Bacci, S., & Ginaldi, M. (2015). A classification of university courses based on students' satisfaction: An application of a two-level mixture item response model. *Quality & Quantity*, 49, 927-940.

- Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture model for multiple choice data. *Journal of Educational and Behavioral Statistics*, 26, 381-409.
- Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, 39, 331-348.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.
- Cho, S.-J. (2007). *A multilevel mixture IRT model for DIF analysis* (Unpublished doctoral dissertation). University of Georgia, Athens, GA.
- Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture model with applications to DIF. *Journal of Educational and Behavioral Statistics*, 35, 336-370.
- Choi, Y.-J. (2014). *Metric identification in mixture IRT models* (Unpublished doctoral dissertation). University of Georgia, Athens, GA.
- Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42, 133-148. doi:10.1111/j.1745-3984.2005.00007
- Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities: Research and Practice*, 20, 225-233.
- Congdon, P. (2003). *Applied Bayesian modelling*. New York, NY: John Wiley.
- Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883-904.
- Fagginger Auer, M. F., Hickendorff, M., van Putten, C. M., Béguin, A. A., & Heiser, W. J. (2016). Multilevel latent class analysis for large-scale educational assessment data: Exploring the relation between the curriculum and students mathematical strategies. *Applied Measurement in Education*, 29, 144-159.
- Finch, W. H., & Finch, M. E. H. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement*, 73, 973-993.
- Foy, P., Arora, A., & Stanco, G. M. (2013). *TIMSS 2011 user guide for the international database*. Chestnut Hill, MA: Boston College.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472. Retrieved from <http://www.jstor.org/stable/2246093>
- Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 202-226.
- Jilke, S., Meuleman, B., & van de Walle, S. (2015). We need to compare, but how? Measurement equivalence in comparative public administration. *Public Administration Review*, 75, 36-48.
- Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, 33, 353-373.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, 41, 499-532.
- Lukociene, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, 40, 247-283.
- Lukociene, O., & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In A. Fink, L. Berthold, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 241-249). Berlin, Germany: Springer.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195-215.

- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535-569.
- Plummer, M., Best, N., Cowles, K., & Vines, K. (2006, March). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6, 7-11.
- Preinerstorfer, D., & Formann, A. K. (2011). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, 65, 251-262.
- Raczynski, K. A., Alagoz, C., Thompson, J., & Pollack, S. (2015). *EngageME P.L.E.A.S.E. Evaluation study results*. Athens, GA: Program Evaluation Group.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: SAGE.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.
- Samuelsen, K. M. (2008). Examining differential item functioning from a latent class perspective. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 67-113). Charlotte, NC: Information Age Publishing.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Sclove, L. S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333-343.
- Sen, S. (2014). *Robustness of mixture IRT models to violations of latent non-normality* (Unpublished doctoral dissertation). University of Georgia, Athens, GA.
- Sen, S., Cohen, A. S., & Kim, S.-H. (2016). The impact of non-normality on extraction of spurious latent classes in mixture IRT models. *Applied Psychological Measurement*, 40, 98-113.
- Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7, 221-242.
- Spiegelhalter, D., Thomas, A., & Best, N. (2003). WinBUGS (Version 1.4) [Computer software]. Cambridge, UK: Biostatistics Unit, Institute of Public Health.
- Tay, L., Diener, E., Drasgow, F., & Vermunt, J. K. (2011). Multilevel mixed-measurement IRT analysis: An explication and application to self-reported emotions across the world. *Organizational Research Methods*, 14, 177-207.
- Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in a growth mixture model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317-341). Charlotte, NC: Information Age Publishing.
- Varriale, R., & Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research*, 47, 247-275.
- Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, 33, 213-239.
- Vermunt, J. K. (2007, August). *Multilevel mixture item response theory models: An application in education testing*. Paper presented at the 56th session of the International Statistical Institute, Lisbon, Portugal.
- Vermunt, J. K. (2008). Multilevel latent variable modeling: An application in education testing. *Austrian Journal of Statistics*, 37, 285-299.
- Vermunt, J. K. (2011). Mixture models for multilevel data sets. In J. Hox & K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 59-81). New York, NY: Routledge.
- Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17, 228-243.
- Wang, A., & Cohen, A. S. (2008, July). *Evaluation of three test speededness models*. Paper presented at the International Meeting of the Psychometric Society, Durham, NH.
- Yang, C. C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, 50, 1090-1104.
- Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, 44, 41-61.