

Chapter 3

Robustness of Mixture IRT Models to Violations of Latent Normality

Sedat Sen, Allan S. Cohen, and Seock-Ho Kim

Abstract Unidimensional item response theory (IRT) models assume that a single model applies to all people in the population. Mixture IRT models can be useful when subpopulations are suspected. The usual mixture IRT model is typically estimated assuming normally distributed latent ability. Research on normal finite mixture models suggests that latent classes potentially can be extracted even in the absence of population heterogeneity if the distribution of the data is nonnormal. Empirical evidence suggests, in fact, that test data may not always be normal. In this study, we examined the sensitivity of mixture IRT models to latent nonnormality. Single-class IRT data sets were generated using different ability distributions and then analyzed with mixture IRT models to determine the impact of these distributions on the extraction of latent classes. Preliminary results suggest that estimation of mixed Rasch models resulted in spurious latent class problems in the data when distributions were bimodal and uniform. Mixture 2PL and mixture 3PL IRT models were found to be more robust to nonnormal latent ability distributions. Two popular information criterion indices, Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), were used to inform model selection. For most conditions, the performance of BIC index was better than the AIC for selection of the correct model.

3.1 Introduction

Item response theory (IRT) models have been designed to describe the relationship between observed item responses and latent variables (Embretson and Reise 2000). The successful applications of standard IRT models depend on several assumptions such as unidimensionality, invariance, local independence, and monotonicity (Reckase 2009). For instance, one set of item characteristic curves (ICCs) can be

S. Sen (✉)
University of Georgia, Athens, GA, USA
e-mail: sedatsen@harran.edu.tr

A.S. Cohen • S.-H. Kim
Harran University, Sanliurfa, Turkey for Sedat Sen and University of Georgia, GA, USA
e-mail: acohen@uga.edu; shkim@uga.edu

used to describe the relationship between item responses and the underlying latent trait by assuming that all individuals come from a single homogeneous population. However, other modeling approaches may be more appropriate when there are subgroups of respondents with different response-trait relationships. Several models have been developed to overcome violations of standard IRT models including multidimensional IRT models (Reckase 2009), multiple group IRT models (Bock and Zimowski 1997), and mixture IRT (MixIRT) models (Rost 1990; Mislevy and Verhelst 1990). MixIRT models, for example, may be more useful when the invariance assumption is violated (von Davier et al. 2007).

The popularity of MixIRT models has increased with the applications of these models to many psychometric issues such as detecting test speededness (Bolt et al. 2002; Wollack et al. 2003; Yamamoto and Everson 1997) and differential item functioning (Cohen and Bolt 2005; Cohen et al. 2005; Samuelsen 2005), identifying different personality styles (von Davier and Rost 1997), and identifying solution strategies (Mislevy and Verhelst 1990; Rost and von Davier 1993), as well as classifying response sets (Rost et al. 1997).

The MixIRT model is based on finite mixture models (Titterton et al. 1985). Finite mixture models are used in a number of latent variable models including latent class analysis (LCA; Clogg 1995), structural equation models (Arminger et al. 1999; Jedidi et al. 1997), growth mixture models (GMMs) (Li et al. 2001), and factor mixture analysis (FMA; Lubke and Muthén 2005). Typically, finite mixture models are used to explain the underlying heterogeneity in the data by allocating this heterogeneity to two or more latent classes. One problem with the application of these models is that the extracted classes may not always reflect the heterogeneity in the population (Bauer and Curran 2003). It may be possible to obtain some extraneous classes as an artifact of misspecification. For instance, Bauer and Curran (2003) demonstrated that nonnormality in the data can lead to identification of spurious latent classes even in the absence of population heterogeneity (McLachlan and Peel 2000; Bauer and Curran 2003). Similar situations have been observed in mixture Rasch models when model specific assumptions are violated (Alexeev et al. 2011).

In contrast to application of multiple group IRT models, the number of groups (or classes) may not be known a priori in exploratory applications of mixture models. In a typical exploratory approach to determine the number of latent classes, several models may be fit to the data. The model with the best fit is often selected based on some statistical criteria (e.g., information criterion indices). Since the extracted classes are latent (i.e., unobserved), one can never be sure about the true number of latent classes. Thus, identification of the correct number of latent classes has become a longstanding and unresolved issue in finite mixture models research. This issue has been studied for a number of latent variable models (Alexeev et al. 2011; Bauer 2007; Bauer and Curran 2003) or model selection statistics (Li et al. 2009; Nylund et al. 2007; Tofghi and Enders 2007; Wall et al. 2012).

Bauer and Curran (2003) examined the effect of nonnormality on the detection of the number of latent classes in GMMs. Data were generated for single-class data sets with normal and nonnormal distributions and then analyzed with one- and two-class

solutions. Results indicated that a one class solution was a better fit for normal data and a two class solution (i.e., a spurious class) was a better fit for nonnormal data. Results further suggested that data with nonnormal distributions may cause over-extraction of latent classes even in a single homogeneous population.

Tofghi and Enders (2007) investigated the performances of nine different fit indices (information criteria and likelihood based statistics) within the context of GMMs. They showed that the sample-size adjusted BIC (SABIC; Sclove 1987) and the Lo–Mendell–Rubin (LMR; Lo et al. 2001) likelihood ratio test are promising in determining the number of classes. Similarly, Nylund et al. (2007) compared the performances of information criteria and hypothesis tests using the likelihood ratio test with three different mixture models: LCA, factor mixture models (FMMs), and GMMs. Results indicated that the bootstrap likelihood ratio test (BLRT) performed better than LMR or likelihood-ratio tests for determining the correct number of classes in the LCA models with continuous outcomes, the FMM and the GMM models. Results also showed that the Bayesian information criterion (BIC; Schwarz 1978) was superior to Akaike’s information criterion (AIC; Akaike 1974) and consistent AIC (CAIC; Bozdogan 1987) for all three types of mixture model analyses. Li et al. (2009) examined the performances of five fit indices for dichotomous mixture Rasch, 2-parameter (2PL), and 3-parameter logistic (3PL) IRT models using an MCMC algorithm. Results of a simulation study showed that in most conditions BIC performed better than the deviance information criterion (Spiegelhalter et al. 1998), AIC, pseudo Bayes factor (PsBF), and posterior predictive model checks (PPMC).

Alexeev et al. (2011) investigated the effects of violation of the Rasch model assumption of equal discriminations on detection of the correct number of latent classes in a mixture Rasch model. Spurious latent classes were observed when data generated with a single-class 2PL IRT model were analyzed with a mixture Rasch model. Results showed further that even a single item with a high discrimination could result in detection of a second class even though the data were generated to be a single class.

Even small departures from model assumptions may have an effect on the number of latent classes detected as well as on model parameter estimates (Alexeev et al. 2011; Bauer 2007; Bauer and Curran 2003). Although latent nonnormality has been investigated in the context of IRT (Bock and Aitkin 1981; Seong 1990; Woods 2004; Zwinderman and Van den Wollenberg 1990), similar work has not been reported with MixIRT models. As was shown for the GMM (Bauer and Curran 2003), it is important to know whether the nonnormality may be responsible for generating additional latent classes in MixIRT models. The purpose of this study, therefore, was to examine the impact of distributional conditions on the extraction of latent classes. We do this in the context of MCMC estimation with dichotomous MixIRT models.

3.2 Method

A Monte Carlo simulation study was conducted to investigate the following research question: Is the accuracy of detection of latent classes affected by using a normal prior on ability parameters when the latent ability distribution is nonnormal?

3.2.1 Simulation Design

The following conditions were simulated: Sample size (600 and 2,000 examinees), test length (10 and 28 items), and five ability distributions (bimodal symmetric, normal, platykurtic, skewed, and uniform). Data were simulated for each of the three dichotomous IRT models \times 3 MixIRT models \times 2 latent class models (LCMs; one- and two-classes) \times 2 sample sizes \times 2 test lengths \times 5 ability distributions = 360 conditions. Twenty-five replications were generated for each condition.

Examinee ability parameters were simulated for normal, skewed, platykurtic, bimodal symmetric, and uniform distributions. For the normal distribution condition, ability parameters were randomly sampled from a standard normal distribution with unit variance (i.e., $N(0, 1)$). Skewed and platykurtic data were generated using the power method proposed by Fleishman (1978). Skewness and kurtosis values were 0.75 and 0.0 for skewed data and 0.0 and -0.75 for platykurtic data, respectively. These values were selected to represent typical nonnormality situations as described by Pearson and Please (1975) for skewness less than 0.8 and kurtosis between -0.6 and 0.6 . For the uniform condition, ability parameters were randomly drawn from $\text{Uniform}(-2, 2)$. The ability parameters for the bimodal symmetric condition were randomly drawn from a combination of two normal distributions: $N(-1.5, 1)$ and $N(1.5, 1)$. All of the conditions were generated using a program written in R (R Development Core Team 2011). Graphical representations of the four nonnormal generating distributions are presented in Fig. 3.1. A standard normal distribution curve is superimposed on each figure for reference. It should be noted that these are actual generating distributions for ability parameters.

Generating item parameters were obtained for the Rasch model, 2PL and 3PL IRT model estimates using data from the Grade 9 mathematics test of the Florida Comprehensive Assessment Test (FCAT; Florida Department of Education 2002) using MULTILOG 7.03 (Thissen 2003). Estimated item parameters for these three models are presented in Tables 3.1 and 3.2 (a —slope parameter, b —threshold parameter, and c —guessing parameter).

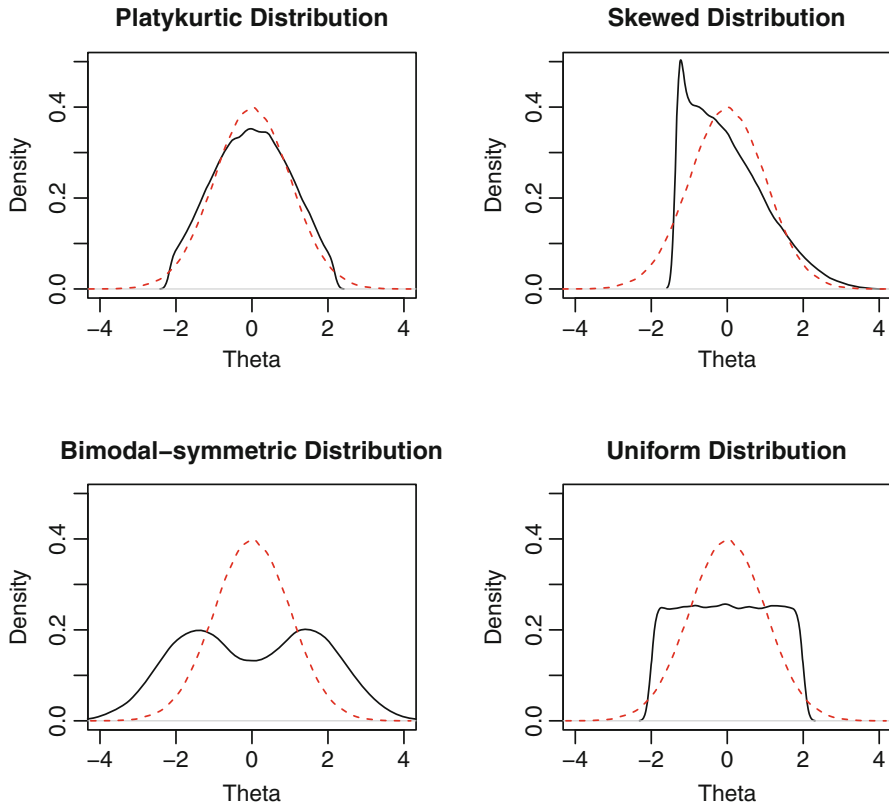


Fig. 3.1 Generating distributions for ability parameters

Table 3.1 Item parameters used for data generation for ten-item condition

Item	Rasch model		2PL model		3PL model	
	b	a	b	a	b	c
1	-1.83	0.91	-1.84	0.91	-0.37	0.53
2	-0.07	0.93	-0.07	1.17	0.69	0.30
3	-0.15	1.21	-0.13	1.23	0.39	0.24
4	0.90	0.84	0.94	0.91	1.23	0.16
5	-0.38	0.94	-0.37	0.66	-0.06	0.12
6	-0.59	1.14	-0.51	0.75	-0.37	0.06
7	0.98	0.76	1.14	0.76	1.38	0.14
8	0.51	1.06	0.45	1.58	0.88	0.22
9	0.99	0.34	2.37	3.87	1.67	0.28
10	0.19	1.27	0.15	1.05	0.46	0.14

Table 3.2 Item parameters used for data generation for 28-item condition

Item	Rasch model		2PL model		3PL model	
	b	a	b	a	b	c
1	-1.72	1.05	-1.66	1.45	-0.45	0.50
2	-0.09	0.88	-0.10	1.96	0.76	0.31
3	-0.16	1.24	-0.16	2.10	0.40	0.24
4	0.81	0.72	1.04	1.62	1.35	0.19
5	-0.37	0.93	-0.39	1.14	0.05	0.16
6	-0.57	1.28	-0.50	1.35	-0.34	0.06
7	0.91	0.72	1.16	1.31	1.40	0.15
8	0.45	1.07	0.42	2.82	0.88	0.22
9	0.91	0.38	2.08	3.97	1.67	0.26
10	0.16	1.27	0.12	1.85	0.48	0.15
11	0.69	0.67	0.95	2.42	1.34	0.25
12	0.42	0.94	0.43	2.26	0.93	0.23
13	0.93	0.69	1.26	3.61	1.35	0.22
14	1.22	0.98	1.24	2.67	1.29	0.14
15	0.31	0.94	0.32	1.66	0.81	0.20
16	1.19	0.92	1.25	2.88	1.30	0.16
17	0.27	1.18	0.23	2.47	0.72	0.22
18	-1.54	1.61	-1.15	1.59	-1.15	0.03
19	-0.39	1.69	-0.32	1.83	-0.15	0.06
20	-0.41	1.46	-0.35	1.77	-0.03	0.14
21	-0.34	1.01	-0.34	1.27	0.12	0.17
22	-0.30	1.22	-0.28	2.84	0.46	0.32
23	0.18	1.87	0.08	2.45	0.30	0.09
24	0.09	0.76	0.13	2.03	0.97	0.32
25	0.10	0.70	0.15	1.01	0.72	0.18
26	-0.31	1.01	-0.31	1.12	-0.09	0.08
27	-0.33	0.91	-0.35	0.93	-0.32	0.00
28	-0.47	1.43	-0.39	1.83	-0.01	0.17

3.2.2 Model Framework

The three dichotomous MixIRT models investigated in this study are described below. These models can be viewed as straightforward extensions of traditional Rasch, 2PL and 3PL IRT models, respectively. First, the mixed Rasch model (MRM; Rost 1990) is described below. This model is a combination of two latent variable models: a Rasch model and a LCM. MRMs explain qualitative differences according to the LCM portion of the model and quantitative differences according to the Rasch model portion of the model. The assumption of local independence holds for the MRM as it does for the LCM and Rasch model. In addition, the MRM assumes

that the observed item response data come from a heterogeneous population that can be subdivided into mutually exclusive and exhaustive latent classes (Rost 1990; von Davier and Rost 2007). The conditional probability of a correct response in the MRM can be defined as

$$P(x_{ij} = 1) = P_{ij} = \sum_{g=1}^G \pi_g \frac{\exp(\theta_{jg} - \beta_{ig})}{1 + \exp(\theta_{jg} - \beta_{ig})}, \quad (3.1)$$

where x_{ij} is the 0/1 response of examinee j to item i (0 = incorrect response, 1 = correct response), π_g is the proportion of examinees for each class, θ_{jg} is the ability of examinee j within latent class g , and β_{ig} denotes difficulty of item i within latent class g . As proposed in Rost (1990), certain constraints on item difficulty parameters and mixing proportions are made for identification purposes so that $\sum_i \beta_{ig} = 0$ and $\sum_g \pi_g = 1$ with $0 < \pi_g < 1$.

The probability of a correct response in a mixture 2PL (Mix2PL) IRT model can be written as

$$P(x_{ij} = 1) = P_{ij} = \sum_{g=1}^G \pi_g \frac{\exp[\alpha_{ig}(\theta_{jg} - \beta_{ig})]}{1 + \exp[\alpha_{ig}(\theta_{jg} - \beta_{ig})]}, \quad (3.2)$$

where α_{ig} denotes the discrimination of item i in class g . In the Mix2PL model, both the item difficulty and item discrimination parameters are permitted to be class-specific. Similarly, the mixture 3PL (Mix3PL) IRT model is assumed to describe unique response propensities for each latent class. This model also allows item guessing parameters to differ in addition to item difficulty and discrimination parameters. As for the MRM and Mix2PL model, each latent class also can have different ability parameters. The probability of a correct response for a Mix3PL model can be described as

$$P(x_{ij} = 1) = P_{ij} = \sum_{g=1}^G \pi_g \left(\gamma_{ig} + (1 - \gamma_{ig}) \frac{\exp[\alpha_{ig}(\theta_{jg} - \beta_{ig})]}{1 + \exp[\alpha_{ig}(\theta_{jg} - \beta_{ig})]} \right), \quad (3.3)$$

where γ_{ig} is guessing parameter for item i in class g . The MixIRT models have been applied in a number of studies (e.g., Cohen and Bolt 2005; Li et al. 2009).

3.2.3 MCMC Specification

As is the case with traditional IRT models, MixIRT models can also be estimated either using MLE or MCMC methods in the Bayesian context. MLE algorithms are applied in several software packages including Latent GOLD (Vermunt and Magidson 2005), mdltm (von Davier 2005), Mplus (Muthén and Muthén 2011), R (psychomix package; Frick et al. 2012), and Winmira (von Davier 2001). MCMC

estimation is possible using the WinBUGS computer software (Spiegelhalter et al. 2003), Mplus and proc MCMC in SAS (v. 9.2; SAS Institute, Cary, NC, USA). MRM estimations can be obtained using any of these software packages. The Mix2PL IRT model can be fit using the Latent GOLD, Mplus and WinBUGS programs, however, only the WinBUGS program has the capability at this time of estimating the Mix3PL IRT model. Thus, the computer software WinBUGS was used in this study for estimating all the models to be studied. In this study, the Rasch model, 2PL and 3PL IRT models were generated to have one class. In order to see whether a two-class solution (i.e., a spurious class situation) will fit where a one-class model was simulated, each MixIRT model was fitted with one- and two-class solutions.

MCMC estimation model specifications are described below including specifications of priors and initial values. In two-group model estimations, 0.5 was used as initial values for the mixing proportions. The starting values for all other parameters were randomly generated using the WinBUGS software. The following prior distributions were used for the MRM:

$$\begin{aligned}\beta_{ig} &\sim \text{Normal}(0, 1), \\ \theta_j &\sim \text{Normal}(\mu(\theta), 1), \\ \mu(\theta)_g &\sim \text{Normal}(0, 1), \\ g_j &\sim \text{Bernoulli}(\pi_1, \pi_2), \\ (\pi_1, \pi_2) &\sim \text{Dirichlet}(.5, .5),\end{aligned}$$

where θ_j represents the ability parameter for examinee j , β_{ig} is the difficulty parameter of item i within class g , and $c_j = \{1, 2\}$ is a class membership parameter. Estimates of the mean and standard deviation for each latent class, μ_g and σ_g , can also be estimated via MCMC. As in Bolt et al. (2002), σ_g was fixed at 1 for both groups. A Dirichlet distribution with 0.5 for each parameter was used as the prior for π_g for the two-group models. In addition, a prior on item discrimination was used in the Mix2PL and Mix3PL models. A prior on guessing parameter was also used in the Mix3PL. These two priors are defined as follows:

$$\begin{aligned}\alpha_{ig} &\sim \text{Normal}(0, 1)I(0,), \\ \gamma_g &\sim \text{Beta}(5, 17).\end{aligned}$$

An appropriate number of burn-in and post burn-in iterations needs to be determined in order to remove the effects of starting values and obtain a stable posterior distribution. Several methods have been proposed to determine the convergence assessment and the number of burn-in iterations. The convergence diagnostics by Gelman and Rubin (1992) and Raftery and Lewis (1992) are currently the most popular methods (Cowles and Carlin 1996). In this study, convergence diagnostics were assessed with these two methods using the R package called convergence diagnosis and output analysis for MCMC (CODA; Plummer et al. 2006). For the MRM conditions, 6,000 burn-in iterations and 6,000 post-burn-in iterations were used based on the diagnostic assessment. For the Mix2PL IRT model conditions,

7,000 burn-in iterations and 7,000 post burn-in iterations were used, and 9,000 burn-in iterations and 9,000 post burn-in iterations were used in all Mix3PL IRT model conditions.

3.2.4 Model Selection

For traditional IRT models, model selection is typically done using likelihood ratio test statistics for nested models and information criterion indices for nonnested models. Since MixIRT models are nonnested models, only information criterion indices can be used to determine the correct number of latent classes. Several information criterion indices have been proposed with different penalization terms on the likelihood function. AIC and BIC indices and their extensions (i.e., SABIC and CAIC) are often used to select the best model from among a set of candidate models based on the smallest value obtained from the same data. In this study, only AIC and BIC indices were used. These two indices are discussed below. AIC can be calculated as

$$\text{AIC} = -2\log L + 2p, \quad (3.4)$$

where L is the likelihood function and p is the number of estimated parameters calculated as follows:

$$p = m * I * j + m * j - 1, \quad (3.5)$$

where m can have values from 1 to 3 for the MRM, Mix2PL, and Mix3PL IRT models, respectively, I denotes the number of items, and j is the number of latent classes. For example, $j = 2$ is used for a two-class MixIRT solution. AIC does not apply any penalty for sample size and tends to select more complex models than BIC (Li et al. 2009). As can be seen below, the BIC index applies a penalty for sample size and for the number of parameters. As a result, BIC selects simpler models than AIC. The BIC has been showed to perform better than AIC for selection of dichotomous MixIRT models (Li et al. 2009; Preinerstorfer and Formann 2011). BIC can be calculated as follows:

$$\text{BIC} = -2\log L + p * \log(N), \quad (3.6)$$

where L is the likelihood of the estimated model with p free parameters and $\log(N)$ is the logarithmic function of the total sample size N . It should be noted that the likelihood values in these equations are based on ML estimation. Since we used MCMC estimation, the likelihood values in these equations were replaced with the posterior mean of the deviance $\overline{D(\xi)}$ as obtained via MCMC estimation (Congdon 2003; Li et al. 2009) where ξ represents all estimated parameters in the model.

3.2.5 Evaluation Criteria

Recovery of item parameters was assessed using root mean square error (RMSE) which is computed as follows:

$$\text{RMSE}_{(\beta_i)} = \sqrt{\frac{\sum_{i=1}^I \sum_{r=1}^R (\beta_i - \hat{\beta}_{ir})^2}{RI}}, \quad (3.7)$$

where β_i and $\hat{\beta}_i$ are generating and estimated item difficulty parameters for item i , respectively. I is the number of items and R is the number of replications. This formula was also used for calculation of the RMSE for item discrimination and item guessing parameters. In order to make an accurate calculation, the estimated parameters were placed on the scale of the generating parameters using the mean/mean approach (Kolen and Brennan 2004). It should be noted that item parameter estimates from one-class mixture IRT solutions were used to calculate the RMSE between the generated single-class IRT data sets. In addition, a percentage of correct detection of simulated latent classes was calculated based on smallest AIC and BIC indices for each condition. The proportion of correct detections for the single-class condition was used as the percentage of correct identification.

3.3 Results

As mentioned earlier, each data set was generated to have one class. The data generated by the Rasch model were fitted with the MRM and the data generated by 2PL and 3PL IRT models were fitted with Mix2PL and Mix3PL IRT models, respectively. These three models were fit with one-class and two-class models using standard normal priors on ability parameters for each simulation condition. The mean RMSE values of item parameters for each condition were calculated and are given in Tables 3.3, 3.4, and 3.5. The proportion of correct positives for the three MixIRT models was calculated based on minimum AIC and BIC between one-class and two-class solutions. For instance, the number of classes for the given data set was defined as correct when the information index for a one-class solution was smaller than that of two-class solution. These proportions are presented in Tables 3.6, 3.7, and 3.8 for each condition. Condition names given in the first column of Tables 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8 include model name, number of items, and number of examinees. For example, the condition Rasch10600 indicates a data condition generated with the Rasch model for ten items and 600 examinees.

Table 3.3 summarizes the mean RMSE values of item difficulty parameters for three MixIRT models. Mean RMSE values of item difficulty parameter for MRMs were found to be less than 0.10 for most of the conditions. RMSE values were around 0.15 in only three of the bimodal data conditions. As shown in Table 3.3, mean RMSE values of the Mix2PL and Mix3PL IRT models were larger than those

Table 3.3 Mean RMSE values of item difficulty parameters over 25 replications

Condition	Bimodal	Normal	Platykurtic	Skewed	Uniform
Rasch10600	0.164	0.093	0.083	0.087	0.095
Rasch28600	0.146	0.089	0.091	0.093	0.095
Rasch102000	0.149	0.077	0.085	0.074	0.088
Rasch282000	0.097	0.051	0.050	0.050	0.057
2PL10600	0.337	0.187	0.196	0.179	0.199
2PL28600	0.280	0.131	0.135	0.133	0.131
2PL102000	0.364	0.111	0.136	0.109	0.161
2PL282000	0.286	0.072	0.072	0.077	0.107
3PL10600	0.777	0.391	0.371	0.363	0.387
3PL28600	0.675	0.204	0.206	0.214	0.290
3PL102000	0.776	0.333	0.341	0.339	0.426
3PL282000	0.617	0.132	0.137	0.183	0.230

Table 3.4 Mean RMSE values of item discrimination parameters over 25 replications

Condition	Bimodal	Normal	Platykurtic	Skewed	Uniform
2PL10600	1.677	0.148	0.144	0.155	0.298
2PL28600	1.524	0.131	0.129	0.138	0.275
2PL102000	1.778	0.086	0.098	0.088	0.357
2PL282000	1.813	0.071	0.069	0.078	0.368
3PL10600	1.220	0.574	0.538	0.515	0.522
3PL28600	1.125	0.730	0.744	0.470	0.545
3PL102000	1.280	0.448	0.503	0.511	0.471
3PL282000	2.176	0.417	0.440	0.452	0.452

for the MRM. Mean RMSE values appear to increase as the complexity of model increases. RMSEs for the Mix2PL IRT model condition with 28 items and 2,000 examinees, however, were less than 0.11 for all except the bimodal symmetric distribution. For the Mix2PL analyses, mean RMSE values seemed to decrease as the number of examinees increases. The mean RMSE values for the bimodal distribution were relatively higher for the Mix3PL IRT model. Mean RMSE values were around 0.30 for normal, platykurtic, skewed, and uniform distributions. These results are consistent with previous simulation studies with MixIRT models (Li et al. 2009).

Mean RMSE values for item discrimination parameter estimates for the Mix2PL and Mix3PL IRT models are presented in Table 3.4. As expected, RMSE values for the Mix2PL and Mix3PL IRT models for the bimodal symmetric distribution were the largest. Those for the uniform distribution were the second largest. Mean RMSE values appeared to be smaller for all of the Mix2PL conditions for the normal, platykurtic, and skewed distributions. Mean RMSE values for the

Table 3.5 Mean RMSE values of item guessing parameters over 25 replications

Condition	Bimodal	Normal	Platykurtic	Skewed	Uniform
3PL10600	0.096	0.089	0.088	0.092	0.089
3PL28600	0.061	0.063	0.058	0.073	0.653
3PL102000	0.092	0.085	0.086	0.093	0.093
3PL282000	0.039	0.047	0.049	0.074	0.048

Table 3.6 The correct positive rates for MRM analyses over 25 replications

Condition	Bimodal		Normal		Platykurtic		Skewed		Uniform	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Rasch10600	0	0	2	25	6	23	0	19	0	2
Rasch102000	1	13	0	19	0	22	0	14	0	0
Rasch28600	0	0	20	25	20	25	15	25	10	25
Rasch282000	3	15	7	25	4	25	1	21	0	0

Mix3PL IRT model condition also decreased as the number of examinees increased, although there was no clear pattern as the number of items increased. Table 3.5 summarizes mean RMSE values for the guessing parameter estimates. For most of the conditions, mean RMSE values appeared to decrease as the number of items and the number of examinees increased. Mean RMSE values for item guessing parameters were relatively lower than those for item difficulty and discrimination parameters. This is because the item guessing parameter estimates are always between zero and one. Thus, the recovery of item guessing parameters is often easier than the recovery of other item parameters, particularly discrimination parameters.

Table 3.6 summarizes the correct positive rates for MRM analyses. As shown in Table 3.6, the BIC index performed well in the MRM analysis under normal, platykurtic, and skewed conditions. However, the proportions of correct positives for the BIC index for the bimodal and uniform conditions were low except in the 28 items and 600 examinees condition. The performance of AIC was lower than BIC for the MRM analyses. AIC did not provide high correct identification rates in the normal distribution conditions. Both AIC and BIC showed good performance in data conditions with 28 items and 600 examinees except for bimodal data. In most of the other simulation conditions, the correct positive rate for AIC index was very low and close to zero.

Table 3.7 presents the correct positive rates for Mix2PL IRT model analyses. For almost all conditions, the correct positive rates of the BIC index were found to be almost perfect except for the skewed data conditions. Although the results of the AIC index in the Mix2PL IRT model analyses provided higher correction rates than that of the MRM analyses, the overall performance of AIC index was worse than BIC results. Correct positive rates for AIC ranged from 0 to 10 in more than half of the conditions. Based on the results from AIC index, latent nonnormality causes spurious latent class in Mix2PL IRT model estimation. However, results based on

Table 3.7 The correct positive rates for Mix2PL analyses over 25 replications

Condition	Bimodal		Normal		Platykurtic		Skewed		Uniform	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
2PL10600	5	25	9	25	7	25	1	16	7	25
2PL102000	6	24	2	22	6	25	0	2	2	17
2PL28600	25	25	25	25	24	25	10	25	19	25
2PL282000	18	25	12	22	21	25	0	10	2	25

Table 3.8 The correct positive rates for Mix3PL analyses over 25 replications

Condition	Bimodal		Normal		Platykurtic		Skewed		Uniform	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
3PL10600	15	25	17	25	13	25	14	25	6	25
3PL102000	2	25	3	24	3	22	23	24	2	17
3PL28600	25	25	25	25	25	25	25	25	25	25
3PL282000	17	25	23	25	16	25	21	25	5	25

the BIC index did not show strong evidence for existence of spurious latent class in Mix2PL IRT model estimation with nonnormal latent distributions.

Table 3.7 presents the correct positive rates for Mix3PL IRT model analyses. In all distribution conditions, BIC supported selection of one class in 100 % of the replications at all three sample size \times two test length conditions. Only the conditions with ten items and 2,000 examinees yielded lower results in terms of the BIC index. The number of correct selections was higher for AIC for the Mix3PL model compared to the previous models. Consistent with the previous results, however, the number of correct selections by AIC was lower than for BIC. Further, AIC had problems with selecting the correct model in most of the uniform data conditions. AIC failed to detect the correct model for the ten items and 2,000 examinees one-class condition. It appears that the Mix3PL IRT models were more robust to latent nonnormality than either the MRM or Mix2PL IRT models based on results for both the AIC and BIC.

3.4 Discussion

The two-class MixIRT model was consistently judged to be a better representation of the data than the one-class model when the data were analyzed with the MRM under both bimodal and uniform data conditions. As expected, MRM analyses of the data with normal and typical nonnormal ability distributions (i.e., skewed and platykurtic) did not show any over-extraction. Both of the indices provided similar results; however, the overall performance of AIC was worse than the BIC.

The results of the Mix2PL and Mix3PL analyses showed similar patterns. For most of the conditions, nonnormality did not appear to lead to over-extraction with

either the Mix2PL or MiX3PL IRT models. These results were not consistent with the results of the MRM analyses. However, the relative performance of fit indices in the Mix2PL and Mix3PL IRT model analyses was consistent with the analyses of MRM in that the AIC selected solutions with two-classes more than BIC. This also was consistent with previous research on model selection that found AIC to select more complex model solutions.

Results suggested that latent nonnormality may be capable of causing extraction of spurious latent classes with the MRM. More complex models, however, such as the Mix2PL and Mix3PL appeared to be more robust to latent nonnormality in that both tended to yield fewer spurious latent class solutions. With respect to the penalty term used in the information indices considered here, the more parameters added to the model, the larger the penalty term. In addition, the performance of the information indices used to determine model fit also may be a function of the underlying distribution of the data. Thus the interpretability of the latent classes in any model selected also needs to be considered in determining model selection. Relying only on statistical criteria may not always yield interpretable solutions. Results in this study suggested that it may be misleading, even under the most ideal conditions, to use the AIC index for identifying number of latent classes. Thus, the solution accepted should be expected to have sufficient support not only from statistical criteria but also from the interpretability of the classes. Further research on the impact of different nonnormal distributions would be helpful, particularly with respect to more extreme skewness and kurtosis conditions that can sometimes arise in highly heterogeneous populations. The skewed and platykurtic data sets in this study were limited to typical nonnormality conditions. It may be useful to investigate the effects of extreme violations of normality on detection of the number of latent classes.

References

- Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Autom Control* 19:716–723. doi:10.1109/TAC.1974.1100705
- Alexeev N, Templin J, Cohen AS (2011) Spurious latent classes in the mixture Rasch model. *J Educ Meas* 48:313–332. doi:10.1111/j.1745-3984.2011.00146.x
- Arminger G, Stein P, Wittenberg J (1999) Mixtures of conditional mean- and covariance-structure models. *Psychometrika* 64:475–494. doi:10.1007/BF02294568
- Bauer DJ (2007) Observations on the use of growth mixture models in psychological research. *Multivar Behav Res* 42:757–786. doi:10.1080/00273170701710338
- Bauer DJ, Curran PJ (2003) Distributional assumptions of growth mixture models: implications for over-extraction of latent trajectory classes. *Psychol Methods* 8:338–363. doi:10.1037/1082-989X.8.3.338
- Bock RD, Aitkin M (1981) Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46:443–459. doi:10.1007/BF02293801
- Bock RD, Zimowski MF (1997) Multiple group IRT. In: van der Linden WJ, Hambleton RK (eds) *Handbook of modern item response theory*. Springer, New York, pp 433–448

- Bolt DM, Cohen AS, Wollack JA (2002) Item parameter estimation under conditions of test speededness: application of a mixture Rasch model with ordinal constraints. *J Educ Meas* 39:331–348. doi:10.1111/j.1745-3984.2002.tb01146.x
- Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52:345–370
- Clogg CC (1995) Latent class models. In: Arminger G, Clogg CC, Sobel ME (eds) *Handbook of statistical modeling for the social and behavioral sciences*. Plenum Press, New York, pp. 311–359
- Cohen AS, Bolt DM (2005) A mixture model analysis of differential item functioning. *J Educ Meas* 42:133–148. doi:10.1111/j.1745-3984.2005.00007
- Cohen AS, Gregg N, Deng M (2005) The role of extended time and item content on a high-stakes mathematics test. *Learn Disabil Res Pract* 20:225–233. doi:10.1111/j.1540-5826.2005.00138.x
- Congdon P (2003) *Applied Bayesian modelling*. Wiley, New York
- Cowles MK, Carlin BP (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *J Am Stat Assoc* 91:883–904. doi:10.1080/01621459.1996.10476956
- Embretson SE, Reise SP (2000) *Item response theory for psychologists*. Erlbaum, Mahwah
- Fleishman AI (1978) A method for simulating non-normal distributions. *Psychometrika* 43:521–532. doi:10.1007/BF02293811
- Florida Department of Education (2002) *Florida Comprehensive Assessment Test*. Tallahassee, FL: Author
- Frick H, Strobl C, Leisch F, Zeileis A (2012) Flexible Rasch mixture models with package psychomix. *J Stat Softw* 48(7):1–25. Retrieved from <http://www.jstatsoft.org/v48/i07/>
- Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Stat Sci* 7:457–472. Retrieved from <http://www.jstor.org/stable/2246093>
- Jedidi K, Jagpal HS, DeSarbo WS (1997) Finite mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Mark Sci* 16:39–59. doi:10.1287/mksc.16.1.39
- Kolen MJ, Brennan RL (2004) *Test equating: methods and practices*, 2nd edn. Springer, New York
- Li F, Duncan TE, Duncan SC (2001) Latent growth modeling of longitudinal data: a finite growth mixture modeling approach. *Struct Equ Model* 8:493–530. doi:10.1207/S15328007SEM0804_01
- Li F, Cohen AS, Kim S-H, Cho S-J (2009) Model selection methods for mixture dichotomous IRT models. *Appl Psychol Meas* 33:353–373. doi:10.1177/0146621608326422
- Lo Y, Mendell NR, Rubin DB (2001) Testing the number of components in a normal mixture. *Biometrika* 88:767–778. doi:10.1093/biomet/88.3.767
- Lubke GH, Muthén BO (2005) Investigating population heterogeneity with factor mixture models. *Psychol Methods* 10:21–39. doi:10.1037/1082-989X.10.1.21
- McLachlan G, Peel D (2000) *Finite mixture models*. Wiley, New York
- Mislevy RJ, Verhelst N (1990) Modeling item responses when different subjects employ different solution strategies. *Psychometrika* 55:195–215. doi:10.1007/BF02295283
- Muthén LK, Muthén BO (2011) *Mplus user's guide*, 6th edn. Author, Los Angeles
- Nylund KL, Asparouhov T, Muthén BO (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct Equ Model* 14:535–569. doi:10.1080/10705510701575396
- Pearson ES, Pleuse NW (1975) Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika* 62:223–241. doi:10.1093/biomet/62.2.223
- Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. *R News* 6:7–11. Retrieved from http://cran.r-project.org/doc/Rnews/Rnews_2006-1.pdf#page=7
- Preinerstorfer D, Formann AK (2011) Parameter recovery and model selection in mixed Rasch models. *Br J Math Stat Psychol* 65:251–262. doi:10.1111/j.2044-8317.2011.02020.x
- R Development Core Team (2011) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. Retrieved from <http://www.R-project.org/>
- Raftery AE, Lewis S (1992) How many iterations in the Gibbs sampler. *Bayesian Stat* 4:763–773

- Reckase MD (2009) *Multidimensional item response theory*. Springer, New York
- Rost J (1990) Rasch models in latent classes: an integration of two approaches to item analysis. *Appl Psychol Meas* 14:271–282. doi:10.1177/014662169001400305
- Rost J, von Davier M (1993) Measuring different traits in different populations with the same items. In: Steyer R, Wender KF, Widaman KF (eds) *Psychometric methodology. Proceedings of the 7th European meeting of the psychometric society in Trier*. Gustav Fischer, Stuttgart, pp 446–450
- Rost J, Carstensen CH, von Davier M (1997) Applying the mixed-Rasch model to personality questionnaires. In: Rost R, Langeheine R (eds) *Applications of latent trait and latent class models in the social sciences*. Waxmann, New York, pp 324–332
- Samuelsen KM (2005) *Examining differential item functioning from a latent class perspective*. Doctoral dissertation, University of Maryland
- SAS Institute (2008) *SAS/STAT 9.2 user's guide*. SAS Institute, Cary
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464. doi:10.1214/aos/1176344136
- Sclove LS (1987) Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52:333–343. doi:10.1007/BF02294360
- Seong TJ (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Appl Psychol Meas* 14:299–311. doi:10.1177/014662169001400307
- Spiegelhalter DJ, Best NG, Carlin BP (1998) Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Research Report No. 98-009. MRC Biostatistics Unit, Cambridge
- Spiegelhalter D, Thomas A, Best N (2003) WinBUGS (version 1.4) [Computer software]. Biostatistics Unit, Institute of Public Health, Cambridge
- Thissen D (2003) MULTILOG: multiple, categorical item analysis and test scoring using item response theory (Version 7.03) [Computer software]. Scientific Software International, Chicago
- Titterton DM, Smith AFM, Makov UE (1985) *Statistical analysis of finite mixture distributions*. Wiley, Chichester
- Tofighi D, Enders CK (2007) Identifying the correct number of classes in a growth mixture model. In: Hancock GR, Samuelsen KM (eds) *Mixture models in latent variable research*. Information Age, Greenwich, pp 317–341
- Vermunt JK, Magidson J (2005) Latent GOLD (Version 4.0) [Computer software]. Statistical Innovations, Inc., Belmont
- von Davier M (2001) WINMIRA 2001 [Computer software]. Assessment Systems Corporation, St. Paul
- von Davier M (2005) mdlm: software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. ETS, Princeton
- von Davier M, Rost J (1997) Self monitoring-A class variable? In: Rost J, Langeheine R (eds) *Applications of latent trait and latent class models in the social sciences*. Waxmann, Muenster, pp 296–305
- von Davier M, Rost J (2007) Mixture distribution item response models. In: Rao CR, Sinharay S (eds) *Handbook of statistics. Psychometrics*, vol 26. Elsevier, Amsterdam, pp 643–661
- von Davier M, Rost J, Carstensen CH (2007) Introduction: extending the Rasch model. In: von Davier M, Carstensen CH (eds) *Multivariate and mixture distribution Rasch models: extensions and applications*. Springer, New York, pp 1–12
- Wall MM, Guo J, Amemiya Y (2012) Mixture factor analysis for approximating a nonnormally distributed continuous latent factor with continuous and dichotomous observed variables. *Multivar Behav Res* 47:276–313. doi:10.1080/00273171.2012.658339
- Wollack JA, Cohen AS, Wells CS (2003) A method for maintaining scale stability in the presence of test speededness. *J Educ Meas* 40:307–330. doi:10.1111/j.1745-3984.2003.tb01149.x
- Woods CM (2004) *Item response theory with estimation of the latent population distribution using spline-based densities*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill

- Yamamoto KY, Everson HT (1997) Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In: Rost J, Langeheine R (eds) Applications of latent trait and latent class models in the social sciences. Waxmann, Munster, pp 89–98
- Zwiderman AH, Van den Wollenberg AL (1990) Robustness of marginal maximum likelihood estimation in the Rasch model. *Appl Psychol Meas* 14:73–81. doi:10.1177/014662169001400107