The Impact of Non-Normality on Extraction of Spurious Latent Classes in Mixture IRT Models

Applied Psychological Measurement
2016, Vol. 40(2) 98–113
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615605080
apm.sagepub.com

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

Abstract

Unidimensional, item response theory (IRT) models assume a single homogeneous population. Mixture IRT (MixIRT) models can be useful when subpopulations are suspected. The usual MixIRT model is typically estimated assuming a 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 non-normal. In this study, the authors examined the sensitivity of MixIRT models to latent non-normality. Single-class IRT data sets were generated using different ability distributions and then analyzed with MixIRT models to determine the impact of these distributions on the extraction of latent classes. Results suggest that estimation of mixed Rasch models resulted in spurious latent class problems in the data when distributions were bimodal and uniform. Mixture two-parameter logistic (2PL) and mixture three-parameter logistic (3PL) IRT models were found to be more robust to latent non-normality.

Keywords

mixture IRT model, latent non-normality, over-extraction, spurious latent class, MCMC estimation

Introduction

The standard, unidimensional item response theory (IRT) models (Lord & Novick, 1968) are known as strong modeling methods because their successful applications require several assumptions such as unidimensionality, invariance, local independence, monotonicity, and existence of a continuous function that describes the relationship between the probability of correct response and latent trait (Reckase, 2009). For instance, the assumption that all examinees are drawn from a single homogeneous population implies that one set of item characteristic curves (ICCs) can be used to describe the relationship between item responses and the underlying latent trait. When the potential exists for subgroups of respondents each with different response–trait

¹Harran University, Şanlıurfa, Turkey

²University of Georgia, Athens, GA, USA

Corresponding Author:

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

Email: sedatsen@harran.edu.tr

relationships, however, it may be useful to consider other modeling approaches. In such a case, mixture IRT (MixIRT) models (Mislevy & Verhelst, 1990; Rost, 1990) may provide a more useful approach for modeling the relationship between test items and latent construct(s).

MixIRT models are based on finite mixture models. Their use has increased in educational and psychological sciences with extensions to several latent variable models (Muthén, 2008). A general use of mixture models is to explain the underlying heterogeneity in the data by allocating it to two or more latent classes. One issue that has arisen with the use of these models, however, is that the extracted classes may not always reflect a heterogeneous population structure (Bauer & Curran, 2003). Rather, it may reflect some extraneous characteristic of the data. Research on finite mixture models suggests that non-normal distributions in the data may produce spurious latent classes even in the absence of population heterogeneity (Bauer & Curran, 2003; McLachlan & Peel, 2000).

When the number of classes within a population is not known a priori, the usual practice is to conduct an exploratory analysis to determine the number of latent classes. This is done by fitting appropriate models with different numbers of classes and then selecting the best fitting among candidate models. Selecting the best fitting model among the alternatives can be done using information-based indices such as Akaike's (1974) information criterion (AIC) or Schwarz's (1978) Bayesian information criterion (BIC). The issue of extracting the correct number of classes has become a long-standing and unresolved issue for researchers who apply finite mixture models. Because of the increase in use of MixIRT models in educational and psychological research, it is important to ensure the detection of the correct number of latent classes in the data. Most of the research on over-extraction within the latent variable modeling context has focused on either the violation of model specific assumptions (Alexeev, Templin, & Cohen, 2011; Bauer & Curran, 2003) or model selection statistics (Li, Cohen, Kim, & Cho, 2009; Nylund, Asparouhov, & Muthén, 2007; Tofighi & Enders, 2007). These studies also differed in the types of models inspected.

Bauer and Curran (2003) and Alexeev et al. (2011) noted that even small departures from model assumptions may have an effect on the number of latent classes detected as well as on model parameter estimates. Although Alexeev et al. demonstrated that violations of the Rasch model (RM) assumption of equal item discriminations (Rasch, 1960) produced spurious classes, no research has yet been reported on the effects of different distributions of the latent variable on the number of classes extracted in MixIRT models. As a significant proportion of test score distributions are likely to be non-normal (Gregoire & Driver, 1987; Micceri, 1989), examining effects of potential non-normality on the extraction of latent classes with MixIRT models is important.

In this study, the impact of distributional conditions was examined on the extraction of latent classes with MixIRT models. Dichotomous item response data (with a correct response coded as 1 and an incorrect response coded as 0) were generated by a single-class IRT model for several different ability distributions and then analyzed with a mixture RM (MixRM), mixture two-parameter logistic model (Mix2PLM), and mixture three-parameter logistic model (Mix3PLM) using estimation methods assuming normal distributions. Descriptions of MixIRT models used in this study are presented below. In addition, issues concerning estimation, model selection, and latent non-normality in MixIRT models are discussed.

MixIRT Models

MixIRT models have been developed by combining an IRT model with a latent class model. The MixIRT models considered in this article assume a unidimensional IRT model for each latent class, and each class has a different set of item and ability parameters. The Mix3PLM is

presented below (see Equation 3).¹ This equation is a partial marginalization from an overall joint likelihood of the data and the latent traits θ_j and g_j . Specifically, this joint distribution can be thought of as having the notation of $f(\mathbf{X}_j, \theta_j, g_j)$, where a person's observed item response data are formed by the row vector $\mathbf{X}_j$. The marginal likelihood of the data is then given by

$$P(\mathbf{X}=\mathbf{x}) = \int \sum_{g=1}^G f(g_j, \theta_j) \left[\prod_{i=1}^I P(x_{ij}=1|\theta_j, g_j)^{x_{ij}} (1 - P(x_{ij}=1|\theta_j, g_j))^{1-x_{ij}} \right] d\theta_j, \quad (1)$$

where the inner terms in Equation 1 are the likelihood of the data $\mathbf{X}$ given a person's ability θ_j and latent class g_j and

$$f(\mathbf{X}=\mathbf{x}|g_j, \theta_j) = \left[\prod_{i=1}^I P(x_{ij}=1|\theta_j, g_j)^{x_{ij}} (1 - P(x_{ij}=1|\theta_j, g_j))^{1-x_{ij}} \right]. \quad (2)$$

In Equation 2, the (prior) likelihood of the joint distribution of ability and latent class, $f(g_j, \theta_j)$, which is the product of conditional and marginal distributions of the latent traits (i.e., $f(g_j, \theta_j) = f(\theta_j|g_j)f(g_j)$).

Thus, the probability of a correct response for a Mix3PLM can be described as

$$P(x_{ij}=1|\theta_j) = P_{ij} = \sum_{g=1}^G \pi_g P(x_{ij}=1|\theta_j, g_j) = \sum_{g=1}^G \pi_g \left(\gamma_{ig} + (1 - \gamma_{ig}) \frac{\exp[\alpha_{ig}(\theta_j - \beta_{ig})]}{1 + \exp[\alpha_{ig}(\theta_j - \beta_{ig})]} \right), \quad (3)$$

where π_g are the mixing proportions, θ_j is the latent ability for person j and α_{ig} denotes the discrimination of item i in class g , β_{ig} indicates the difficulty of item i within latent class g , and γ_{ig} is the guessing parameter for item i in class g .

The Mix2PLM and MixRM are nested within the Mix3PLM in Equation 3. For example, the conditional probability of a correct response for the Mix2PLM can be obtained by setting the guessing parameter to zero. The MixRM model conditional probability can be obtained by constraining both item discrimination and item guessing parameters. To ensure identification for the MixRM, certain constraints on item difficulty parameters and mixing proportions should be made. Rost (1990) has proposed that $\sum_i \beta_{ig} = 0$. Furthermore, because the model is a discrete mixture model, $\sum_g \pi_g = 1$ with $0 < \pi_g < 1$. The probability of belonging to each of the latent classes can also be estimated for every individual using the posterior probability; individuals are then assigned to the latent class with the highest probability.

In traditional IRT models, parameter estimation can be done using maximum likelihood estimation (MLE) or Bayesian estimation (e.g., Markov chain Monte Carlo [MCMC]). Parameter estimation of MixIRT models can be more complex than of traditional IRT models due to the increase in the number of parameters to be estimated. Parameter estimation for MixIRT models can also be applied using either MLE or MCMC methods. Bayesian estimation of MixRM requires the specification of prior distributions for all parameters to be estimated. In the context of the MixRM, estimates of ability parameters, θ_j , item difficulty, β_{ig} , group membership, g , and mixture probabilities, π_g , are of interest. For the Mix2PLM, α_i , an item discrimination parameter is also estimated, and for the Mix3PLM, a guessing parameter, γ_i , is also estimated. A normal prior is often used for the distributions of ability and item difficulty. Cho, Cohen, and Kim (2013) noted that use of normal priors has been found to yield reasonable estimates of the item parameters in MixIRT models. Group membership has a multinomial distribution composed of one draw on g categories with mixture probabilities (e.g., $g \sim \text{Multinomial}(1, \pi_g)$;

Congdon, 2003; McLachlan & Peel, 2000). A Dirichlet prior and a Dirichlet process with stick-breaking prior are described by McLachlan and Peel (2000) as conjugate priors for mixture probabilities. Both forms of Dirichlet prior are conjugate to a multinomial distribution. A stick-breaking process is typically used to generate a Dirichlet process by iteratively breaking the proportions from the remainder with a unit-length stick. The MCMC algorithm is used to sample a class membership parameter for each examinee at each stage of the Markov chain (Bolt, Cohen, & Wollack, 2002; Cho et al., 2013). For instance, in a two-class MixRM, class membership parameters, $g = 1, 2$, are sampled for each examinee j . Then, a group membership parameter for each examinee is determined as the mode of g across iterations following burn-in (Cho et al., 2013).

Model selection. Determining the number of classes in the data is another important issue in the use of MixIRT models. In general, likelihood ratio (LR) test statistics or information criterion indices can be used for model selection within the IRT context. Under MLE, the LR test provides a uniformly most powerful test. As Bayesian estimation was used in the present study, information criterion indices appropriate for Bayesian estimation were used to inform model selection. AIC and BIC are two commonly reported indices for model selection. Both were initially developed for use in the context of MLE. Congdon (2003) has described analogues of these based on the posterior of the likelihood (see below). In addition, Spiegelhalter, Best, Carlin, and van der Linde (2002) described the deviance information criterion (DIC) as a generalization of AIC for use in the context of Bayesian estimation. Kang and Cohen (2007) reported DIC to perform less well than either AIC or BIC for model selection of standard IRT models. Li et al. (2009) likewise found both AIC and DIC to be less accurate than BIC for model selection of MixIRT models. For this reason, BIC as described in Congdon was used for model selection in this study. Results for AIC are also reported but for comparison with previous research.

These indices are based on some form of penalization of the likelihood function. Typically, information indices are used to compare different model solutions of the same data when the maximum likelihood (ML) estimates of the parameters are obtained. Smaller values of information criterion indices indicate better fit. However, they may provide different solutions to the same data due to differences in the penalty function applied to likelihood. Two of the more widely used information criteria, AIC and BIC, are discussed below. AIC can be calculated as follows:

$$\text{AIC} = -2 \log L + 2d, \quad (4)$$

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

$$d = m \times I \times j + 2 \times j - 1, \quad (5)$$

where m can have values from 1 to 3 for the MixRM, Mix2PLM, and Mix3PLM, 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. As shown in Equation 4, $2d$ is used as a penalty for over parameterization in the AIC index. One problem with AIC is that it does not take the sample size into account and also tends to select more complex models (Li et al., 2009). The lack of penalization for sample size leads to inconsistency in performance of AIC (Tofighi & Enders, 2007). However, BIC applies a penalty function that uses both the number of parameters and the sample size. BIC has been found to be somewhat more accurate than AIC for selection of MixIRT models (Li et al., 2009; Preinerstorfer & Formann, 2011). BIC tends to select simpler models than AIC due to the inclusion of the number of parameters in the penalty function. BIC can be calculated as

$$\text{BIC} = -2 \log L + d \times \ln(N). \quad (6)$$

AIC and BIC, as described above, are based on a likelihood estimated using MLE. Use of Bayesian estimation requires a different likelihood function, specifically the posterior mean of the deviance. The MLE based deviance value was replaced in this study with the posterior mean of the deviance $\overline{D(\xi)}$ as obtained via MCMC estimation (Congdon, 2003) where ξ represents all estimated parameters in the model. The likelihood function was monitored at each iteration, and the final likelihood value was taken as the posterior mean of the deviance. Li et al. (2009) reported that this form of BIC was more accurate than either AIC or DIC with MixIRT models.

Latent non-normality. The distribution of the latent variable is typically assumed to be normal in most IRT applications, particularly when unidimensional IRT models are fitted using marginal maximum likelihood estimation (MMLE). The distribution of the latent variable should be checked before applying unidimensional IRT, as erroneously assuming the latent variable to be normally distributed may have a biasing effect on parameter estimates (Woods, 2004). Several methods have been proposed for dealing with latent non-normality in IRT, including alternative parametric forms (Andersen & Madsen, 1977), the empirical histogram method (Bock & Aitkin, 1981), using Johnson curves (Thissen, 1991), and using Ramsay-curve IRT (RC-IRT; Woods, 2004). The impact of latent non-normality on parameter estimates, however, has not been well studied in MixIRT models. Xu and Jia (2011) estimated MixRM and Mix2PLM with generalized skewed normal and standard normal distributions. Item parameter estimates and ability distributions were similar in all simulated conditions, but standard errors of item parameters were found to be larger in the skewed distribution condition.

Two simulation studies are reported below to investigate the accuracy of detection of latent classes affected by using a normal prior on ability parameters when the latent ability distribution is non-normal and the effects of skewness and kurtosis on the extraction of latent classes for MixIRT models. Finally, an empirical study is used to illustrate how one might detect and handle latent non-normality.

Simulation Study I

A simulation study was conducted to examine the impact of non-normality on the accuracy of detection of latent classes. In Study 1, data were simulated for a single population but with different latent distributions.

Methods for Simulation Study I

Simulation conditions. Two test lengths were simulated, a short length 10-item test and a medium length 28-item test. In addition, two sample sizes were simulated, a small sample, 600-examinee condition and a large sample, 2,000-examinee condition. These conditions are similar to those reported in Li et al. (2009). As noted in Li et al., samples of 600 appeared to be appropriate for one- to four-group MixRMs and possibly for a Mix2PLM for 15- and 30-item tests. Rost (1990) analyzed a MixRM with a sample size of 1,800, and Samuelsen (2005) used a sample size of 2,000 examinees.

Four distributions were simulated: normal, platykurtic, skewed, uniform, and bimodal. These distributions have been reported in previous research on latent non-normality (see, for example, Woods, 2004).

Three MixIRT models were fit to these data, each model estimating one or two latent classes. There was a total of 360 conditions: 2 sample sizes (600 and 2,000) $\times$ 2 test lengths (10 and 28 items) $\times$ 3 dichotomous IRT models (Rasch, 2PL, and 3PL) $\times$ 5 ability distributions (normal, platykurtic, skewed, uniform, and bimodal) $\times$ fitting of 3 MixIRT models (MixRM, Mix2PLM, and Mix3PLM) $\times$ fitting one- and two-class MixIRT models. Fifty replications were simulated for each condition.

Skewed and platykurtic data were generated using Fleishman's (1978) power method. Skewed data were randomly drawn from a distribution with generating values of 0.75 and 0.0, and platykurtic data were randomly drawn from a distribution with values of 0.0 and -0.75 . These represent typical non-normality situations in which distributions have skewness less than 0.8 and kurtosis between -0.6 and 0.6 (e.g., Pearson & Please, 1975). Uniform data were randomly drawn from a uniform $(-2, 2)$ distribution. Ability parameters for the bimodal symmetric condition were randomly drawn from a combination of two normal distributions $N(-1.5, 1)$ and $N(1.5, 1)$. Finally, data for the normal distribution simulation were randomly drawn from a standard normal distribution $N(0, 1)$. This condition was added for completeness as it is frequently used for estimating IRT models. Graphs of the four non-normal conditions are given in Figure A1 (see the online appendix). A normal distribution curve is superimposed on each for reference. Generating item parameters for Study 1 were obtained from Rasch, 2PL, and 3PL model estimates using data from a Grade 9 mathematics test for a large southeastern state.

Estimation of models. A MCMC algorithm was used as implemented in the computer software WinBUGS (Spiegelhalter, Thomas, & Best, 2003) for estimation of models in this study. When the target densities are log-concave, as expected in these analyses, the adaptive rejection sampling method is automatically implemented in the WinBUGS software (Patz & Junker, 1999). For the mixing proportions, 0.5 was used as Dirichlet weights. This treats both groups as equally weighted. The starting values for all other parameters were randomly generated using the WinBUGS software. The posterior mean of the likelihood was used to compute AIC and BIC values. The following priors and hyper-priors were used for estimation of the MixIRT models: $\beta_{ig} \sim N(0, 1)$, $\theta_j \sim N(\mu(\theta)_g, 1)$, $\mu(\theta)_g \sim N(0, 1)$, $g_j \sim \text{Bernoulli}(\pi_1, \pi_2)$, $(\pi_1, \pi_2) \sim \text{Dirichlet}(.5, .5)$, $\alpha_{ig} \sim N(0, 1) I(0, \infty)$, $\gamma_{ig} \sim \text{Beta}(5, 17)$; where θ_j represents the ability parameter for examinee j , β_{ig} is the difficulty parameter of item i within class g , π_1 and π_2 are class membership parameters, μ_g is the mean for class g , and $I(0, \infty)$ designates the positive range restriction. Following Bolt et al. (2002), σ_g^2 , the variance for group g , was fixed at 1. A Dirichlet distribution was used as the prior for π_g for the two-group models.

Subscript g can be dropped for the one-group analysis. Gelman–Rubin (Gelman & Rubin, 1992) convergence statistics were used to determine burn-in and post-burn-in iterations. A conservative number of burn-in iterations was selected for analyses of the three different MixIRT models. For the MixRM, 6,000 iterations were discarded as burn-in, and 6,000 were used as post-burn-in iterations. For the Mix2PLM, 7,000 burn-in iterations and 7,000 post-burn-in iterations were used. For the Mix3PLM, 9,000 burn-in iterations and 9,000 post-burn-in iterations were used.

Results for Simulation Study I

Dimensionality assessment. The dimensionality of each generated data set in each condition was first analyzed as a manipulation check to ensure it was unidimensional before including it in the study data set. A three-step process was followed. First, each data set was screened for non-significant chi-square statistics to test the difference between the adequate fit of a one-factor model (null hypothesis) and adequate fit of a two-factor model (alternative hypothesis). Data

sets that had a non-significant chi-square were retained as they were considered unidimensional. Second, data sets that had a significant chi-square were re-examined to determine the proportion of variance accounted for. The data sets with a dominant factor of 20% proportion of variance were assumed to be unidimensional using Reckase's (1979) criterion. Third, data sets that did not meet either of these two criteria were dropped, and new data sets were generated until 50 unidimensional data sets were generated for each condition. The data generated by the RM were fitted with the MixRM. Likewise, the data that were generated by 2PL and 3PL models were fitted with Mix2PLM and Mix3PLM, respectively. These models were fit with a one-class solution and a two-class solution using standard normal priors on ability parameters for each simulation condition.

Recovery analysis for Study 1. Recovery of item parameter estimates was assessed using root mean square errors (RMSEs) calculated between the generating parameters and the parameter estimates as indicated below. Estimates for item difficulty and discrimination were placed on the same scale as the generating parameters using the mean-mean linking method. Lower RMSE values indicate better estimation accuracy. The computational formula used for RMSE for difficulty parameters is

$$\text{RMSE}_{(\beta_i)} = \sqrt{\frac{\sum_{i=1}^I \sum_{r=1}^R (\beta_i - \hat{\beta}_{ir})^2}{RI}}, \quad (7)$$

where β_i and $\hat{\beta}_{ir}$ are generating and estimated difficulty parameters for item i , respectively; I ($i = 1, \dots, I$) indicates the number of items; and R ($r = 1, \dots, R$) is the number of replications. Similar RMSEs were calculated for discrimination parameters. Posterior estimates for AIC and BIC were calculated for each replication and used to determine whether the one-class or the two-class models were the better fit to the generated data. The number of detections of the simulated model was calculated based on AIC and BIC for each condition. As only a one-class model was used to generate data, the number of correct detections was defined as the number of times the one-class model was selected.

Item parameter estimates from the one-class solution were used for recovery analyses because the one-class solution is the MixIRT equivalent of the single-class IRT models that were used for data generation. Mean RMSE values for item parameter estimates of Simulation Study 1 are presented in Table A1 (see the online appendix). For the difficulty parameter, RMSEs of less than 0.10 were obtained for estimates from the MixRM (see Table A1). These were the smallest for the three models in Study 1. The accuracy of the item difficulty parameter appeared to decrease as the model complexity increased. For the Mix2PLM analyses, mean RMSE values ranged from 0.07 to 0.34; Mix3PLM analyses resulted in mean RMSE values between 0.15 and 0.60. The distribution condition also appeared to have an effect on the recovery of the item difficulty parameter. The Mix2PLM conditions with bimodal symmetric distributions yielded the highest RMSE values, ranging from 0.29 to 0.34. The rest of the Mix2PLM conditions yielded mean RMSE values around 0.15. The mean RMSE values for difficulty parameter estimates from Mix3PLM conditions with the bimodal symmetric distribution were relatively large and ranged from 0.56 to 0.60. Mean RMSEs for the Mix3PLM for the uniform distribution were also relatively large, ranging from 0.26 to 0.35. These RMSEs indicate that recovery of item difficulty for the Mix3PLM was less accurate in the bimodal symmetric and uniform conditions in Study 1.

Table 1. Number of Correct Detections Over 50 Replications for Study 1.

Condition	Bimodal		Normal		Platykurtic		Skewed		Uniform	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
RM10600	0	0	6	49	10	47	3	41	0	3
RM102000	0	0	0	39	0	40	0	22	0	0
RM28600	0	0	42	50	41	50	33	50	19	49
RM282000	0	0	35	50	31	50	1	50	0	0
2PL10600	12	50	16	46	15	48	1	29	15	50
2PL102000	18	49	2	40	7	46	0	6	2	32
2PL28600	48	50	49	50	47	50	20	49	34	50
2PL282000	34	50	47	50	45	50	0	24	2	50
3PL10600	46	50	10	50	9	50	14	50	7	48
3PL102000	18	50	1	36	2	33	3	46	0	44
3PL28600	50	50	49	50	48	50	50	50	48	50
3PL282000	46	50	42	50	14	50	17	50	7	50

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; RM = Rasch model; 2PL = two-parameter logistic; 3PL = three-parameter logistic.

For the discrimination parameter, conditions with bimodal symmetric distributions tended to produce estimates with larger RMSEs than those with other distributions. Conditions with uniform distributions appeared to have the second most biased estimates in the Simulation Study 1.

For the platykurtic, skewed, and uniform distribution conditions, mean RMSE values for item discrimination ranged from 0.07 to 0.37 for the Mix2PLM and from 0.20 to 0.60 for the Mix3PLM. Bimodal symmetric distribution conditions yielded mean RMSE values that were greater than 1.0. Mean RMSEs for item discrimination under the bimodal symmetric distribution ranged between 1.5 and 1.8 for the Mix2PLM and between 1.5 and 2.0 for the Mix3PLM.

Mean RMSE values for the guessing parameter estimates ranged from 0.04 to 0.11 for all four distribution conditions.

Number of correct detections for Simulation Study 1. AIC and BIC values were calculated for each MCMC run. AIC is reported for comparison with previous research; BIC was used as the primary information index for model selection. The fit indices were used to determine whether a two-class model had better fit than a one-class model. The number of correct detections was calculated as the number of detections of the one-class (i.e., the generating) model over 50 iterations.

These frequencies are reported in Table 1. The abbreviations for condition names indicate the condition under which the data were generated. For example, the abbreviation RM102000 indicates that the data were generated with a RM for 10 items and 2,000 examinees.

As shown in the first four rows of Table 1, the number of correct detections for the MixRM for the bimodal symmetric and uniform conditions were all either zero or very low except for the uniform distribution condition with 28 items and 600 examinees. The performance of the BIC index was better for the smaller sample size and the longer test length conditions. The performance of AIC was poorer than BIC for the MixRM for all conditions under each of the five distributions.

Type of distribution appeared to have less of an effect on correct model detection for the Mix2PLM. The number of correct detections presented in the middle four rows of Table 1 indicates that for most conditions, BIC correctly detected close to 50 out of the 50 replications. The

number of correct detections was lower for the large sample conditions for the 10-item tests. Correct detections were generally poor for most of the skewed distribution conditions.

The AIC index provided higher correct detection numbers for the Mix2PLM than were observed for the same index for the MixRM. Overall performance of AIC index was worse than the BIC results in Mix2PLM analyses. Consistent with the results for BIC, correct detections for AIC were generally lower in the 10-item conditions than in the 28-item conditions. Furthermore, these decreased as the number of examinees increased. Based on the results from AIC index, latent non-normality appears to cause detection of spurious latent classes in the Mix2PLM model. Results based on the BIC index, however, were more accurate for this model in these conditions.

For the Mix3PLM under all five distribution conditions, BIC supported selection of the correct one-class model in all of the 50 replications except for the 10-item and 2,000-examinee condition. The conditions with the normal and skewed distributions yielded the lowest BIC detection results. Based on the results with BIC, latent non-normality seemed to be a problem for the MixRM in the bimodal symmetric and uniform distributions. Latent non-normality appeared to be less of a problem for the Mix2PLM and Mix3PLM. The Mix3PLM appeared to be somewhat more robust to latent non-normality than either the MixRM or Mix2PLM.

Simulation Study 2

Results from Study 1 suggested that the latent non-normality present in the form of skewness and kurtosis may be factors in the extraction of spurious latent classes. The purpose of Study 2 was to investigate this issue further by examining the impact of different amounts of skewness or kurtosis.

Methods for Simulation Study 2

Simulation conditions. The power method proposed by Fleishman (1978) was again used to generate six different combinations of skewness and kurtosis. Graphs of the six distributions are presented in Figure A2 (see the online appendix). A standard normal distribution curve was superimposed on each of the graphs for reference purposes.

As was the case in Study 1, two sample sizes (600 and 2,000 examinees) were simulated for two test lengths (10 and 28 items). The generating values for item parameters in Study 2 were the same as used in Study 1. In summary, the following factors were simulated for Study 2: 3 IRT models for data generation $\times$ 3 MixIRT models for estimation $\times$ 2 classes for fitting (one class and two classes) $\times$ 2 sample sizes $\times$ 2 test lengths $\times$ 6 ability distributions = 432 conditions. Fifty replications were simulated for Study 2.

The generating models for Study 2 were the one-class models used in Study 1. The same generating values for item parameters were used for Study 2 as for Study 1.

Results for Simulation Study 2

Recovery analysis for Study 2. Mean RMSE values for each of the conditions are given in Table A2 (see the online appendix). RMSEs for the MixRM were the lowest of the three models, ranging from 0.055 to 0.146 for estimates of item difficulties for the MixRM. It also was more difficult to recover item difficulties in Condition 1 (see column headed COND1 in Table A2) than in the other conditions for all three MixIRT models. The increase in kurtosis appeared to increase the RMSEs for item difficulties. The effects of skewness, however, did not show any clear

Table 2. Number of Correct Detections Over 50 Replications.

Condition	Condition 1		Condition 2		Condition 3		Condition 4		Condition 5		Condition 6	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
RM10600	6	42	1	40	2	36	4	42	6	41	9	47
RM102000	0	15	0	3	0	2	0	31	0	35	2	35
RM28600	40	50	20	50	14	50	41	50	43	50	45	49
RM282000	0	50	0	50	0	45	2	50	12	50	25	50
2PL10600	3	30	2	29	0	20	9	40	16	45	12	45
2PL102000	0	2	0	1	0	0	0	17	6	42	12	44
2PL28600	19	50	10	50	1	50	42	50	47	50	47	50
2PL282000	0	29	0	33	0	26	6	48	29	49	45	50
3PL10600	21	50	31	50	33	50	15	48	15	50	12	49
3PL102000	3	48	5	47	5	46	1	45	4	46	1	37
3PL28600	46	50	49	50	50	50	50	50	49	50	50	50
3PL282000	32	50	34	50	22	50	46	50	48	50	43	50

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; RM = Rasch model; 2PL = two-parameter logistic; 3PL = three-parameter logistic.

pattern. The mean RMSEs for item difficulty estimates for the Mix2PLM ranged from 0.09 to 0.25 and increased for the Mix3PLM (see Table A2). For the Mix3PLM, RMSEs ranged between 0.21 and 0.49 (see Table A2).

The mean RMSE values for item discrimination parameter estimates ranged from 0.10 to 0.26 for Mix2PLM estimates and from 0.21 to 1.04 for Mix3PLM estimates. Item discrimination parameter estimates for the Mix3PLM appeared to be more biased than those from the Mix2PLM in the Study 2. Conditions with the highest kurtosis and skewness values (Conditions 1, 2, and 3) yielded the highest RMSEs for item discrimination. Mix3PLM estimates with 28 items and 2,000 examinees under the distribution Condition 3 were recovered least well with a mean RMSE value of 1.04. Almost all of the simulation conditions in Study 2 yielded smaller mean RMSE (<0.164) values for guessing. Data conditions with the highest skewness values (Conditions 1, 2, and 3) appeared are the ones with the highest RMSEs.

Number of correct detections for Simulation Study 2. As was the case for Study 1, the number of correct detections for the three MixIRT models was calculated based on minimum AIC and BIC between the one-class and two-class solutions. These frequencies are presented in Table 2 for each condition. Condition names are given in the first column of Table 2 and indicate the condition used to generate the data. For example, the condition 2PL282000 indicates data that were generated with the 2PL IRT model for 28 items and 2,000 examinees.

The correct detections for the MixRM analyses are presented in the first four rows of Table 2. BIC detected the correct model for most MixRM estimates for the 28-item data sets. The correct detections using the BIC index were low for the 10-item and 2,000-examinee data sets. The performance of the BIC index appeared to be lower under the first three conditions than the performances under the last three conditions. For Condition 2, it was 3 out of 50 correct, and for Condition 3, it was 2 out of 50 correct for the 10-item and 2,000-examinee data sets. The performance of the AIC index was lower than BIC index for the MixRM analyses. Based on AIC, detection of spurious latent classes appeared to be a problem for conditions with the extreme skewness and kurtosis values (Conditions 1, 2, 3, and 4). Correct detections using AIC were generally higher only in the 28-item and 600-examinee conditions. The correct rates for the AIC index appeared to increase as the number of examinees decreased, and the sample size increased.

Table 2 also presents rates for Mix2PLM for the six conditions. BIC correctly identified the generating model in all of the 28-item and 600-examinee data sets. The correct detections for BIC were lower, however, for the Mix2PLM under the first three distribution conditions. The number of correct detections using BIC was higher for the remaining three conditions. The correct detections based on BIC appeared to increase as the kurtosis decreased (i.e., from Condition 1 to Condition 3). No clear pattern existed with regard to increase or decrease of skewness across the last three conditions. AIC was less effective at selection of the correct model except for the 28-item and 600-examinee condition. The correct detections for the AIC were lower for the first three conditions. Apparently, the increase in kurtosis led to detection of spurious latent classes. The correct detections for AIC for the Mix2PLM analyses appeared to increase as the number of items increased and as the sample size decreased. Overall performance of the AIC index was lower than that of BIC index for Mix2PLM analyses. Latent non-normality appeared to cause extraction of spurious latent classes based on AIC. The results based on BIC showed fewer spurious latent class detections for the Mix2PLM. Only the first three data conditions (Condition 1 to Condition 3) with 10 items and 2,000 examinees resulted in the selection of two-class solutions based on the BIC index.

The correct detections for the Mix3PLM analyses are presented in the lower part of Table 2. BIC supported selection of the correct model for almost all distribution conditions. BIC selected the correct model for more than 90% of the replications in all conditions except Condition 6, with skewness of 0 and kurtosis of 3.5. BIC results for Condition 6 with 10 items and 2,000 examinees were the lowest (37 out of 50). BIC was correct for the conditions with 28 items. For the 10-item conditions, the correct detections for BIC decreased as the sample size increased. The number of correct selections was higher for AIC for the Mix3PLM compared with AIC for the MixRM and Mix2PLM. Consistent with previous results, however, the number of correct selections by AIC was lower than for BIC in all conditions. Consistent with the results of MixRM and Mix2PLM, performance of the AIC index was lower than performance of the BIC index. The first three distribution conditions appeared to yield more spurious latent class extraction problems than the last three distribution conditions. This was not as much of a problem for the Mix3PLM as it appeared to be more robust to latent non-normality than either the MixRM or Mix2PLM.

Empirical Study

An empirical data set was analyzed to illustrate how non-normality can affect results for large-scale assessments. MixIRT models were applied to an eighth-grade mathematics data set collected as part of the 2011 Trends in International Mathematics and Science Study (TIMSS; Foy, Arora, & Stanco, 2013). One of the highest performing countries, South Korea (henceforth, Korea), was selected for the analyses in the empirical study, because it provided a data set with a clearly negatively skewed empirical distribution. Data from answers to the 18 multiple-choice items in Booklet 5 were selected for the MixIRT model analyses. There were 369 examinees who had taken Booklet 5.

RC-IRT analysis was used to examine the normality of the latent ability distribution as the negative skew of the empirical distribution does not necessarily mean that the latent ability distribution is also distributed non-normal. The RC-IRT method estimates the latent density as a B-spline-based density (Woods, 2004) simultaneously with estimation of item parameters. In this study, the RC-IRT method was used as implemented in the computer program RCLOG version 2 (Woods, 2006b). RCLOG is designed to detect and correct non-normality in the data set. It has been shown to perform well for estimation of IRT models for data sets with non-normal latent ability distributions (Woods, 2006a). Item and ability parameters are estimated

through a combination of Bock–Aitkin IRT with Ramsay curves (Woods, 2004). The RCLOG program produces a series of Ramsay curves for the latent distribution based on different number of breaks (two to six) and orders (two to six). Breaks refer to the locations on the horizontal axis where different polynomials are joined together, whereas order indicates the order of the polynomial B-splines (Woods, 2006a). As the number of breaks and order increase, the number of parameters and the flexibility of the density increase (Woods, 2006a).

Empirical Data Results

A chi-square difference test between a one- and a two-factor exploratory factor analysis yielded a result of 33.28 with $df = 17$ ($p < .003$), indicating the hypothesis of non-normality was supported in the data. The proportion of variance accounted for suggested the first factor explained 41.7% (i.e., an eigenvalue of 7.52 out of 18) of the variation in the data set. Based on Reckase's (1979) 20% criterion, this data set was determined to have a dominant first factor structure and, therefore, to be unidimensional.

The normality of the ability distribution of the Korean sample data was examined by fitting Ramsay-curve models with different combinations of skewness and kurtosis to the data. The standard 3PL model was fit to responses to the 18 items on the Korea TIMSS data set. This model was used because it is the model that was used for the TIMSS math test. RC-IRT models with number of breaks and orders from two to six were fit using the RCLOG software. Combining the five breaks and five orders resulted in 25 different models. As recommended in the examples in the RCLOG manual, 121 quadrature points were specified ranging from -6 to 6 , and the value of 75 was used for the standard deviation. Results of fit statistics for each break–order combination were examined for the 25 different Ramsay-curve models. The Ramsay curve with three breaks and order of two had the smallest AIC and BIC values, and therefore the best fit to the data. The skewness and kurtosis of this latent distribution were -1.21 and 2.53 , respectively.

MixIRT Results

The same priors were used as described for Study 1. As was used in the two simulation studies, a binomial prior was used for group membership in the models with two latent classes. A multinomial distribution was used for the prior on group membership for MixIRT models with more than two latent classes (Cho et al., 2013). As was done for Studies 1 and 2, 6,000, 7,000, and 9,000 burn-in and post-burn-in iterations were used for the MixRM, Mix2PLM, and Mix3PLM, respectively. Each of the three models was estimated with from one to five latent classes. Results for AIC and BIC are reported in Table 3. As can be seen in this table, AIC selected the three-class solution and BIC selected the two-class solution for the MixRM. Both AIC and BIC selected the one-class solution for the Mix2PLM and Mix3PLM. As was the case with the two simulation studies and consistent with previous research, AIC tended to select the more complex model resulting in overestimating the number of latent classes. Thus, the two-class solution was identified as the best fit for the MixRM and the one-class solution as the best fit for the Mix2PLM and Mix3PLM based on the BIC index.

Discussion

This study investigated the effect of non-normal latent ability on the number of latent classes extracted by MixIRT models. Two simulation studies were conducted: Study 1 investigated the effects of different types of non-normal distributions, and Study 2 investigated the effects of

Table 3. MCMC-Based Fit Statistics for MixIRT Analyses of Korea TIMSS Data.

Model	MixRM		Mix2PLM		Mix3PLM	
	AIC	BIC	AIC	BIC	AIC	BIC
1-Class	5,237.00	5,312.00	5,074.00	5,219.00	5,058.00	5,273.00
2-Class	5,027.00	5,179.00	5,085.00	5,378.00	5,155.00	5,590.00
3-Class	5,015.00	5,246.00	5,075.00	5,516.00	5,248.00	5,901.00
4-Class	5,046.00	5,355.00	5,220.00	5,810.00	5,369.00	6,242.00
5-Class	5,082.00	5,469.00	5,292.00	6,031.00	5,415.00	6,506.00

Note. MCMC = Markov chain Monte Carlo; MixIRT = mixture IRT; TIMSS = Trends in International Mathematics and Science Study; AIC = Akaike information criterion; BIC = Bayesian information criterion. Minimum AIC and BIC values are presented in bold text for each analysis.

different amounts of kurtosis and skewness on detection of latent classes. Test length and sample size were also manipulated. An empirical study was presented illustrating how the effects of latent non-normality might be detected in real data.

For the MixRM analyses conducted in Simulation Study 1, a two-class model was consistently found to be a better representation of the data than a one-class model for both bimodal and uniform data conditions. The MixRM analyses of the data for normal, skewed, and platykurtic distributions, however, yielded fewer over-extraction problems using the BIC index for model selection. The numbers of correct selections for these three non-normal distributions were low based on the AIC index. Thus, the overall performance of AIC appeared to be less effective in that it tended to select the more complex model in the presence of latent non-normality than did BIC. This appears to be consistent with previous research (e.g., Li et al., 2009) suggesting that AIC tends to select the more complex model, in this case, the model with more latent classes.

The results for the Mix2PLM and Mix3PLM showed similar patterns in terms of relative performance of the AIC and BIC information indices. AIC selected models with two classes more often than did BIC. This is consistent with previous research, in which AIC tended to select the more complex model. For most of the simulated conditions, however, latent non-normality did not appear to result in over-extraction for either the Mix2PLM or Mix3PLM. These models appear to be more robust than the MixRM to violations of latent normality as simulated in this study.

In Study 1, the sample size of 600 had higher numbers of correct detections than the 2,000 sample size, mainly in the 10-item condition. This result for the normal distribution conditions disagrees with results reported by Li et al. (2009). The cause of this disagreement is unclear as the same code as used by Li et al. for estimating both AIC and BIC was available for this study. The Li et al. results, however, were obtained on normally distributed data. It would be useful to further examine this disagreement, for example, by comparing results for different generating models using both MLE and Bayesian estimation.

The results for Simulation Study 2 showed similar patterns to those of Simulation Study 1. Performance of the AIC index was poorer than BIC, and recovery declined as model complexity increased as shown by increases in RMSEs. In addition, the more complex Mix3PLM yielded fewer spurious latent classes than the simpler MixRM. Latent non-normality may be causing extraction of spurious latent classes to some extent for the Mix2PLM. Results for both AIC and BIC suggest that over-extraction is likely for the MixRM.

The results of the empirical study also showed some similarities to results from the two simulation studies. The Korean data set from the 2011 TIMSS eighth-grade mathematics assessment appeared to have a non-normal, somewhat skewed latent ability distribution. When analyzed with the three MixIRT models, the simpler model, the MixRM, extracted two latent classes. Both the Mix2PLM and Mix3PLM, however, only found a single class. As with the two simulation studies, the more complex MixIRT models appeared to be more robust than the MixRM to violations of latent non-normality.

Interpretability of latent classes in any MixIRT model needs to be a consideration in determining model selection as 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 (Bauer & Curran, 2003).

Finally, results of this study suggest that latent non-normality not only caused over-extraction but also resulted in poor estimation of the item parameters. This suggests that, when dealing with empirical data, as opposed to simulated data, it is advisable to test the distribution of ability for the possibility of non-normality. The RC-IRT method provides a useful tool for this kind of analysis. Final model selection also should be interpreted cautiously when severe non-normality is present. Several methods have been proposed for correcting non-normality in the traditional IRT analyses (Bock & Aitkin, 1981; Woods, 2004). Results of this study provide information about how this problem can be handled in the context of MixIRT model estimation.

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.

Note

1. The authors wish to thank an associate editor for helping clarify and improve the discussion for Equations 1 and 2.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.
- Alexeev, N., Templin, J., & Cohen, A. S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, 48, 313-332.
- Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika*, 42, 357-374.
- Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for over-extraction of latent trajectory classes. *Psychological Methods*, 8, 338-363.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- 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.

- Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2013). Markov chain Monte Carlo estimation of a mixture item response model. *Journal of Statistical Computation and Simulation*, 83, 278-306.
- Congdon, P. (2003). *Applied Bayesian modelling*. New York, NY: Wiley.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532.
- Foy, P., Arora, A., & Stanco, G. M. (2013). *TIMSS 2011 user guide for the international database*. Chestnut Hill, MA: Boston College.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- Gregoire, T. G., & Driver, B. L. (1987). Analysis of ordinal data to detect population differences. *Psychological Bulletin*, 101, 159-165.
- Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, 31, 331-358.
- 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.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195-215.
- Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1-24). Charlotte, NC: Information Age.
- 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.
- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24, 146-178.
- Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika*, 62, 223-241.
- 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.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- 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. (2005). *Examining differential item functioning from a latent class perspective* (Doctoral dissertation). University of Maryland, College Park.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Spiegelhalter, D., Thomas, A., & Best, N. (2003). WinBUGS (Version 1.4) [Computer software]. Cambridge, UK: Biostatistics Unit, Institute of Public Health.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583-639.
- Thissen, D. (1991). *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory*. Chicago, IL: Scientific Software International.
- Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in a growth mixture model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Mixture models in latent variable research* (pp. 317-341). Greenwich, CT: Information Age.

- Woods, C. M. (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.
- Woods, C. M. (2006a). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, 11, 253-270.
- Woods, C. M. (2006b). *RCLOG v.2: Software for item response theory parameter estimation with the latent population distribution represented using spline-based densities* (Technical report). St. Louis, MO: Washington University.
- Xu, X., & Jia, Y. (2011). *The sensitivity of parameter estimates to the latent ability distribution* (ETS RR-11-40). Retrieved from <http://www.ets.org/Media/Research/pdf/RR-11-40.pdf>