Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 | Article 31

6-5-2020

Comparison of Scale Identification Methods in Mixture IRT Models

Youn-Jeng Choi
University of Alabama, ychoi26@ua.edu

Allan S. Cohen
University of Georgia, acohen@uga.edu

Recommended Citation

Choi, Youn-Jeng and Cohen, Allan S. (2020) "Comparison of Scale Identification Methods in Mixture IRT Models," *Journal of Modern Applied Statistical Methods*: Vol. 18 : Iss. 1 , Article 31.
DOI: 10.22237/jmasm/1556669700
Available at: https://digitalcommons.wayne.edu/jmasm/vol18/iss1/31

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP2971
doi: 10.22237/jmasm/1556669700

בס"ד
Copyright © 2019 JMASM, Inc.
ISSN 1538 − 9472

Comparison of Scale Identification Methods in Mixture IRT Models

Youn-Jeng Choi
University of Alabama
Tuscaloosa, AL

Allan S. Cohen
University of Georgia
Athens, GA

The effects of three scale identification constraints in mixture IRT models were studied. A simulation study found no constraint effect on the mixture Rasch and mixture 2PL models, but the item anchoring constraint was the only one that worked well on selecting correct model with the mixture 3PL model.

Keywords: Item response theory, mixture models, scale identification, MCMC, Bayesian estimation

Introduction

When the same IRT model does not fit all members of a population, a mixture IRT model (MixIRTM) may be appropriate. The MixIRTM is formed by an integration of an IRT model with a latent class model (Mislevy & Verhelst, 1990; Rost, 1990). The IRT part of the model estimates a continuous latent variable and the latent class part estimates a categorical latent variable. Combining these two models permits examining the possibilities that a population of examinees may be classified into some number of discrete latent classes, and that item and ability parameters may differ for each class (Bolt, Cohen, & Wollack, 2002).

Characterizing members of different latent classes is important for interpreting the meaning of the classes. Comparison of item parameter estimates between latent classes is one approach for characterizing the latent classes (Rost, 1990). In order to make such comparisons, however, the latent classes need to have a common metric.

Three methods have been proposed for developing a common metric between latent classes. These three methods are also commonly used in general IRT to fix the metric: item anchoring, person centering, and item centering (de Ayala, 2009).

doi: 10.22237/jmasm/1556669700 | Accepted: September 30, 2018; Published: June 5, 2020.
Correspondence: Youn-Jeng Choi, ychoi26@ua.edu

The first (item anchoring) is concurrent calibration in which one or more items are used to anchor the metrics between classes (Bolt et al., 2002; Choi, Alexeev, & Cohen, 2015; von Davier & Yamamoto, 2004). The second method, person centering, is to impose equality constraints by fixing the mean and standard deviation of one latent class to some values such as zero and one (Baker & Kim, 2004; Cho, Cohen, & Kim, 2013; Cho, Cohen, & Templin, 2008; De Boeck, Cho, & Wilson, 2011). A third method, item centering, is setting the sum of item difficulties to zero for each latent class (Cho & Cohen, 2010; Dai & Mislevy, 2006; Rost, 1990).

Item anchoring may be used when there are either theoretical or empirical reasons for fixing some set of items to given values. If item parameters are known, for example, it is possible to fix the item parameters at known values in each group. When multiple groups are analyzed, these items may be used as anchors to link the metric across groups. As an example, in the likelihood ratio test for differential item functioning (DIF), all item parameter estimates can be constrained to the same values in each group except those of the studied item(s) (Thissen, Steinberg, & Wainer, 1993). Then the item parameters of the studied item(s) are estimated in each group.

Person centering is to impose equality constraints for some reference class by setting the mean of one group to zero and the unit of scale (i.e., its standard deviation) to one. The item and ability parameter estimates for the other groups are then estimated relative to the estimates for the reference group. Person centering is used in programs such as LOGIST (Wingersky, Barton, & Lord, 1982), BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), MULTILOG (Thissen, Chen, & Bock, 2003), and PARSCALE (Muraki & Bock, 2003).

Item centering sets the mean of the item difficulty parameters to zero during calibration. Programs such as WINSTEPS (Linacre, 2001a), BIGSTEPS (Linacre & Wright, 2001), FACETS (Linacre, 2001b), and WINMIRA (von Davier, 2001) use item centering (de Ayala, 2009). Programs such as M-plus (Muthén & Muthén, 2012) and OpenBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007) allow researchers to impose any of these three methods. Although each of the three methods has been reported in the literature, relatively little research exists investigating the impact of these constraints on developing a common metric in MixIRT models.

Rost (1990) described a constraint for the mixture Rasch model (MixRM) in which the mean item difficulty was set to zero. There is somewhat less agreement, however, an about constraint used for the mixture 2PL model (Mix2PLM) or mixture 3PL model (Mix3PLM). Results from Choi, Alexeev, Cohen, and Kim

(2010) for the Mix3PLM indicated that both item centering and person centering worked well for recovery of generating parameters. These results were based on only five replications, however, and for a relatively small number of conditions. Therefore, it is difficult to generalize their conclusion.

The purpose of this study was to investigate the effects of these three constraints on establishing a common metric between latent classes in MixIRT models. An empirical example is provided for motivation and a simulation study is provided examining the impact for each method on selection of the correct model, recovery of item and latent class parameter estimates, and selection of the correct latent class for each examinee.

Mixture IRT Models

MixIRTMs assume there may be groups or classes of examinees that are latent in the population and for which the same IRT model does not hold. An examinee population is assumed to be composed of a fixed number of discrete latent classes each of which fit an IRT model (Bolt, Cohen, & Wollack, 2001). All examinees who belong to a latent class are assumed to be homogeneous on a set of unique characteristics that differentiates one class from another. These models may be appropriate when a single IRT model is not the best fit to the data.

A 3-parameter logistic model is assumed to hold for each class in a Mix3PLM. Item and ability parameters are allowed to differ between latent classes. Each examinee is parameterized both by a class membership parameter ($g = 1, ..., G$) and an ability parameter (θ_j).

The probability of a correct response in the Mix3PLM can be written as

$$P\left(y_{ij} = 1 \mid \theta_j\right) = \sum\nolimits_{g=1}^{G} \pi_g \left[c_{ig} + \left(1 - c_{ig}\right) \frac{\exp\left\{a_{ig}\left(\theta_j - b_{ig}\right)\right\}}{1 + \exp\left\{a_{ig}\left(\theta_j - b_{ig}\right)\right\}} \right], \tag{1}$$

where g indexes latent class ($g = 1, ..., G$), j is the jth examinee among N examinees ($j = 1, ..., N$ examinees), θ_j is the latent ability of examinee j, π_g is the proportion of examinees for each class, a_{ig} is the discrimination parameter for item i in class g, b_{ig} is the difficulty parameter for item i in class g, and c_{ig} is the lower asymptote parameter for item i in class g. Class membership decides the relative difficulty of the items for an examinee in that class. The MixRM and Mix2PLM can be considered as nested within the Mix3PLM in Equation 1 (Sen, Cohen & Kim, 2016). One can obtain the probability of a correct response for the Mix2PLM, for example, by constraining the guessing parameter to zero. Similarly, one can obtain the probability of a correct response for the MixRM by constraining the discrimination parameter to one and the guessing parameter to zero.

Empirical Example: TIMSS 2011 Grade 8 Science Test

An example is provided to illustrate the problem of comparing metrics between different latent classes in the same model. For this example, data were taken from the TIMSS 2011 Grade 8 Science Test (Foy, Arora, & Stanco, 2013; IEA, 2013). Seventeen multiple-choice items and eight short answer items (scored dichotomously) were analyzed for this example. The multiple-choice items were scored correct or incorrect and blanks were skipped and not scored. Approximately .04% (= 2,694/(2,493*25)) of total responses were blanks and ranged from .00% for Item 1 and to 31.13% for Item 24. The short answer items also were scored as either correct or incorrect, and blank items were skipped. The items measured four content domains: Biology (8 items), Chemistry (6 items), Physics (4 items), and Earth Science (7 items).

Sample. Data from seven of the 45 countries in the TIMSS 2011 program were used for this example. The sample of 2,493 students in this data set was from the following countries: Chinese Taipei (*N* = 357, Mean = 564), Ghana (*N* = 410, Mean = 306), Morocco (*N* = 464, Mean = 376), Norway (*N* = 247, Mean = 494), Singapore (*N* = 423, Mean = 590), the Republic of Korea (*N* = 361, Mean = 590), and Ukraine (*N* = 231, Mean = 501). The seven countries were selected, because, as a group, their average scale scores on the test approximated high, middle and low achievement among the participating countries. Singapore, Chinese Taipei, and the Republic of Korea had the highest mean mathematics scores, Ukraine and Norway were average, and Morocco and Ghana were among the lowest for participating countries.

Estimation. The MixRM and Mix2PLM were estimated with each of the three constraints for establishing a common metric: Item anchoring (Constraint 1) was done by using a single anchor item. Person centering (Constraint 2) was done by setting the mean ability of the first latent class to zero with unit variance. Item centering (Constraint 3) was done by setting the mean of item difficulties to zero in each class.

Estimation of model parameters was done using the Markov Chain Monte Carlo (MCMC) algorithm as implemented in the OpenBUGS computer software

(Spiegelhalter et al., 2007). The following conjugate priors were used in the estimation of the MixRM and Mix2PLM in the empirical example: $a_{ig} \sim \text{Normal}(0,1)$ and $a_{ig} > 0$, $i = 1, ..., n$ items; $b_{ig} \sim \text{Normal}(0,1)$; $c_{ig} \sim \text{Beta}(5,17)$; $\theta_j \sim \text{Normal}(\mu_g,1)$, $j = 1, ..., N$ examinees; $\mu_g \sim \text{Normal}(0,1)$, $g = 1, ..., G$ latent classes; and $(\pi_1, ..., \pi_G) \sim \text{Dirichlet}(0.5, ..., 0.5)$; where a is the discrimination parameter, b is the difficulty parameter, c is the lower asymptote parameter (c was used for Mix3PLM in the simulation study). The conjugate priors have been used for default priors for BILOG program and Li, Cohen, Kim, & Cho (2009) also used same prior information. All these priors also were used in the simulation study.

Heidelberger and Welch (1983) convergence diagnostics were used to determine the number of iterations as implemented in the Coda package using R (Plummer, et al., 2012). Autocorrelations, density plots, and history plots were also examined for further evidence of convergence. For the MixRM, a burn-in of 8,000 iterations and 22,000 post-burn-in iterations were found to be sufficient for convergence for all parameters with Constraint 1 (item anchoring) and Constraint 3 (item centering). A burn-in of 2,000 iterations and post-burn-in iterations of 24,000 were sufficient for convergence for Constraint 2 (person centering). For the Mix2PLMs, a burn-in of 8,000 iterations and 21,000 post-burn-in iterations were sufficient for Constraint 1. A burn-in of 9,000 iterations and 16,000 post-burn-in iterations were used for Constraint 2, and a burn-in of 3,000 iterations and 27,000 iterations for Constraint 3 were used. The Mix3PLM did not appear to be converging for all three constraint conditions. Therefore, it was not appropriate to include the Mix3PLM in this empirical study.

Results for Example

Model Selection. BIC (Schwartz, 1978) was used to inform model selection for the MixRM and Mix2PLM as described by Congdon (2003). AIC values (Akaike, 1973) were calculated for comparison purposes, as described by Congdon (2003), although results from Li et al. (2009) suggest BIC may be more accurate for MixIRT models. Smaller AIC and BIC values indicate the better fitting model.

Both indices suggested different numbers of latent classes. AIC suggested five latent classes using Constraint 1 (item anchoring), six latent classes using Constraint 2 (person centering), and seven latent classes using Constraint 3 for the MixRM (item centering). BIC suggested five latent classes using Constraints 1 and 2 and six classes using Constraint 3. For the Mix2PLM, AIC suggested four latent classes for the Mix2PLM using Constraints 1 and 2 and a 3-class solution using

Constraint 3. Based on BIC, a 3-class solution was suggested for all three constraints (see Table 1).

Label Switching. Label switching can occur in both maximum likelihood estimation and MCMC estimation. It can be observed in real data when latent classes switch during a single MCMC chain. It also can be inferred when multiple modes exist of the posterior densities for class membership. In addition, if different latent classes for a MixIRT model have higher percentages of agreement under the different constraints, then a second type of label switching may be inferred. For this example, labels were switched based on the highest percentages of agreement for group membership.

Table 1. Model Comparison Information Criteria for MixRMs and Mix2PLMs

		AIC			**BIC**		
	Latent Classes	**Constraint 1 (Item Anchoring)**	**Constraint 2 (Person Centering)**	**Constraint 3 (Item Centering)**	**Constraint 1 (Item Anchoring)**	**Constraint 2 (Person Centering)**	**Constraint 3 (Item Centering)**
MixRMs	1	67670	67670	67670	67820	67820	67820
	2	65890	65870	65870	66190	66170	66180
	3	65210	65180	65200	65660	65640	65660
	4	64910	64810	64850	65510	65430	65470
	5	**64520**	64460	64490	**65280**	**65240**	65270
	6	64520	**64420**	64240	65430	65350	**65170**
	7			**64190**			65280
Mix2PLMs	1	66430	66460	66460	66580	66750	66760
	2	65300	65550	65210	65890	66140	65810
	3	64660	64550	**64570**	**65540**	**65450**	**65470**
	4	**64370**	**64330**	64570	65540	65530	65770

An example of agreement results for class membership between different constraints using BIC for the MixRM is shown in Table 2. Values on the main diagonal indicate the number of exact agreements: 959 examinees (38.5%) of the sample were placed into Class 1 by both Constraints 1 and 2. The percent matching was 91.2% between Constraints 1 (item anchoring) and 2 (person centering) and 94.3% between Constraints 1 (item anchoring) and 3 (item centering). There was no label switching observed for the MixRMs with Constraints 1 and 2 but label switching was observed for the MixRM between Constraints 1 and 3.

Table 2. Latent Class Classifications for the MixRM with Constraint 1 and Constraint 2

MixRM with Constraint 1	**MixRM with Constraint 2**					
	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**	***Total***
Class 1	959	31	0	0	3	*993*
	(38.5%)	(1.2%)	(0.0%)	(0.0%)	(0.1%)	*(39.8%)*
Class 2	1	239	1	0	0	*241*
	(0.0%)	(9.6%)	(0.0%)	(0.0%)	(0.0%)	*(9.7%)*
Class 3	132	6	990	1	0	*1129*
	(5.3%)	(0.2%)	(39.7%)	(0.0%)	(0.0%)	*(45.3%)*
Class 4	20	14	0	48	3	*85*
	(0.8%)	(0.6%)	(0.0%)	(1.9%)	(0.1%)	*(3.4%)*
Class 5	6	0	1	0	38	*45*
	(0.2%)	(0.0%)	(0.0%)	(0.0%)	(1.5%)	*(1.8%)*
Total	*1118*	*290*	*992*	*49*	*44*	*2493*
	(44.8%)	*(11.6%)*	*(39.8%)*	*(2.0%)*	*(1.8%)*	*(100.0%)*

The percentage of agreement for classification of class membership for the Mix2PLM was 80.8% between Constraints 1 and 2 and 89% between Constraints 1 and 3. Label switching was observed for the Mix2PLMs between Constraints 2 and 3.

Comparison of Class Means and Latent Class Proportions. Additional equating or scale transformation was not required for comparisons of scale parameters within each constraint as this was accomplished by each of the constraints. Comparisons of scale parameters between constraints for the same model, however, did require an additional scale transformation (Choi et al., 2010). Mean and sigma equating was used for these transformations. Means for each of latent classes are reported in Table 3 and indicate differences among the three constraints for the MixRMs. As an example, the differences between means for Constraints 2 and 3 appear to be relatively large for classes 1, 2, 4, and 5. For the Mix2PLM, the means for class 2 using constraint 2 also differed from those for the other two constraints.

The proportions of examinees in latent classes 1 to 6 for each constraint are reported in Table 4. Those for the MixRM look somewhat similar although the proportions for the first and third classes differ for each of the three constraints. The Mix2PLMs had different proportions of class membership for each constraint. The proportions in class 3 look similar for the three constraints, however, the membership proportions of classes 1 and 2 look different for the different constraints. Joint classifications for the MixRM and Mix2PLM also differed for each of the constraints. These results clearly suggest different classifications for the different constraints.

Table 3. Latent Class Means for MixRM and Mix2PLM

	Mixture Rasch Model			Mixture 2PL Model		
Latent Classes	Constraint 1	Constraint 2	Constraint 3	Constraint 1	Constraint 2	Constraint 3
1	0.38	0.00	0.78	-1.24	-1.43	-1.58
2	0.77	0.59	1.12	1.34	-0.07	0.86
3	-1.18	-1.47	-0.72	0.90	0.92	0.79
4	0.15	-0.50	0.46			
5	1.45	0.90	1.46			
6			-0.24			

Table 4. Proportions of Latent Classes for MixRM and Mix2PLM

	Mixture Rasch Model			Mixture 2PL Model		
Latent Classes	Constraint 1	Constraint 2	Constraint 3	Constraint 1	Constraint 2	Constraint 3
1	39.8	44.8	43.8	47.6	31.4	38.9
2	9.7	11.6	9.5	29.5	47.2	38.7
3	45.3	39.8	40.7	22.9	21.5	22.4
4	3.4	2.0	3.9			
5	1.8	1.8	1.8			
6			0.3			

Correlations between parameter estimates for each MixIRTM under each of the constraints should be high if constraints had no impact. Most correlations between constraints in each latent class for the MixRM were high (*r* = .99). Those between Constraints 1 and 2 (*r* = .918) and between Constraint 2 and 3 (*r* = .928) in Class 4, however, were slightly smaller, suggesting that there might be some effect of constraints in Class 4. Correlations between discrimination parameter estimates within latent classes ranged from .88 to .99 in the Mix2PLM. Correlations between difficulty estimates in each latent class suggested estimates differed between constraints in the Mix2PLM. Exceptions were correlations between Constraints 1 and 3 in Class 1 (*r* = .98) and between Constraints 1 and 2 in Class 3 (*r* = .99).

Conclusions. For both models, the constraints had a somewhat different effect on item difficulty estimates, ability estimates, numbers of latent classes, classifications of examinees into latent classes, and proportions of membership in each latent class. The number of latent classes extracted differed for each constraint used with the MixRM, but not for the Mix2PLM. Mean ability, proportions of group memberships, and item parameter estimates also differed depending on the constraint used.

Simulation Study

A simulation study was used to better understand the impact of the three constraints in the context of three MixIRT models: MixRM, Mix2PLM, and Mix3PLM. Simulation conditions included two sample sizes (600 examinees and 2,400 examinees), two test lengths (20 and 40 items), and three different cases of latent classes (1-, 2-, and 3-classes with different proportions of simulated classes for each of these three MixIRT models. Simulated proportions of 30% and 70% were used for the two latent class simulations and for the three latent class simulations 60%, 30%, and 10% were used. Twenty replications were done of the three constraints × two test lengths × two sample sizes × one to three latent classes × three mixture IRT models = 108 conditions.

MCMC estimation was used for model estimation using the same priors as in the example. The Heidelberger and Welch (1983) convergence diagnostic was used to monitor convergence. The number of latent classes was determined using the Bayesian information criterion (BIC). AIC was also monitored. At each iteration, the posterior mean of the deviance was used to calculate AIC and BIC.

Table 5. Simulated Performance Patterns

Type of Knowledge	Group 1	Group 2	Group 3
1	Good	Average	Poor
2	Average	Poor	Good
3	Poor	Good	Average

Three types of knowledge were simulated in each test as suggested by Li et al. (2009). The generating parameters for the knowledge type are given in Table 5. Generating parameters for the MixIRTM are given in Table 6. Those for Items 1 to 5 were the same for the three latent classes and were used as anchors for Constraint 1. 25% of items were designed as anchor items following the suggestion in Kolen and Brennan (2004) that the anchors should be at least 1/5 of the total test in length.

Item parameters for the remaining items were used to simulate three types of knowledge. Items 6 to 10 simulated Type 1 knowledge, Items 11 to 15 simulated Type 2 knowledge, and Items 16 to 20 simulated Type 3 knowledge.

Table 6. Generating parameters for MixIRT Model Simulations: 25% Anchor Items

		Class 1			Class 2			Class 3		
Type of Knowledge	**Item**	**b**	**a**	**c**	**b**	**a**	**c**	**b**	**a**	**c**
1, 2 & 3 (Anchors)	1	-0.50	1	0.20	-0.50	1	0.20	-0.50	1	0.20
	2	-0.50	1	0.20	-0.50	1	0.20	-0.50	1	0.20
	3	0.00	1	0.20	0.00	1	0.20	0.00	1	0.20
	4	0.50	1	0.20	0.50	1	0.20	0.50	1	0.20
	5	0.50	1	0.20	0.50	1	0.20	0.50	1	0.20
1	6	-2.00	2	0.10	-0.50	1	0.20	1.00	1	0.30
	7	-1.75	2	0.10	-0.25	1	0.20	1.25	1	0.30
	8	-1.50	2	0.10	0.00	1	0.20	1.50	1	0.30
	9	-1.25	2	0.10	0.25	1	0.20	1.75	1	0.30
	10	-1.00	2	0.10	0.50	1	0.20	2.00	1	0.30
2	11	-0.50	1	0.20	1.00	1	0.30	-2.00	2	0.10
	12	-0.25	1	0.20	1.25	1	0.30	-1.75	2	0.10
	13	0.00	1	0.20	1.50	1	0.30	-1.50	2	0.10
	14	0.25	1	0.20	1.75	1	0.30	-1.25	2	0.10
	15	0.50	1	0.20	2.00	1	0.30	-1.00	2	0.10
3	16	1.00	1	0.25	-2.00	2	0.10	-0.50	1	0.20
	17	1.25	1	0.25	-1.75	2	0.10	-0.25	1	0.20
	18	1.50	1	0.25	-1.50	2	0.10	0.00	1	0.20
	19	1.75	1	0.25	-1.25	2	0.10	0.25	1	0.20
	20	2.00	1	0.25	-1.00	2	0.10	0.50	1	0.20

Note: a = discrimination, b = difficulty, and c = lower asymptote parameters

Class 1 was simulated to have good performance on Type 1 knowledge, average performance on Type 2 knowledge, and poor performance on Type 3 knowledge. Class 2 was simulated to have average, poor and good performance on Types 1, 2, and 3 knowledge, respectively. Class 3 was simulated to have poor, good and average performance on Types 1, 2, and 3 knowledge, respectively. The item parameters for class 1 were used for the 1-class model. Item parameters for classes 1 and 2 were used for the 2-class model. The 3-class model was simulated using the item parameters for classes 1, 2, and 3. The pattern for the 20-item test was used twice for the 40-item test.

Recovery Evaluation. Bias, root mean square error (RMSE), and Pearson correlations were calculated to evaluate the accuracy of the estimates of item and class mean parameters. RMSE was computed as the square root of the average of the square of the distance between the estimator and its generating parameter (see Equation 5). The bias, RMSE, and Pearson correlation were computed across items, latent class groups, and replications by Equations 4 and 6.

$$Bias\left(\hat{b}\right)=\frac{\Sigma_{r=1}^{R}\Sigma_{g=1}^{G}\Sigma_{i=1}^{I}\left(\hat{b}_{igr}-b_{ig}\right)}{RGI}, \tag{2}$$

$$RMSE\left(\hat{b}\right)=\sqrt{\frac{\Sigma_{r=1}^{R}\Sigma_{g=1}^{G}\Sigma_{i=1}^{I}\left(\hat{b}_{igr}-b_{ig}\right)^{2}}{RGI}}, \tag{3}$$

$$Cor\left(\hat{b},b\right)=\frac{1}{R}\Sigma_{r=1}^{R}\frac{Cov\left(\hat{b}_{igr}-b_{ig}\right)}{\sigma_{\hat{b}_{igr}}\sigma_{b_{ig}}} \tag{4}$$

where $\hat{b}_{igr}$ is the estimated item difficulty parameter for item *i* in latent group *g* for *r*th replication, b_{ig} is the generating true value of item difficulty for item *i* in latent group *g*, *R* is the number of replications (*r* = 1, ..., *R*), *I* (*I* = 1, ..., *I*) is the number of items, and *G* (*g* = 1, ..., *G*) is the number of latent classes in the model being estimated.

In order to estimate bias and RMSE, item parameter estimates need to be on the same scale as the generating parameters. The metrics of estimates from each replication were transformed to the metric of generating parameters using the mean and sigma equating method (de Ayala, 2009; Kolen & Brennan, 2004).

Monitoring Convergence. Three convergence diagnostics were used: the Heidelberger and Welch (1983), the ratio of the standard deviation of the parameter estimate to the MC standard error for the parameter estimate, and the 95% credibility interval. For the MixRM and Mix2PLM, chains were monitored using Heidelberger and Welch convergence diagnostics and the ratio of the standard deviation to the MC standard error for the parameter. The 95% credibility interval was used to monitor convergence of the Mix3PLM.

The chain for the MixRM was found to have converged for all parameters after a burn-in of 5,000 iterations and 5,000 post-burn-in iterations. The chain for the Mix2PLM converged with a burn-in of 6,000 iterations and a post burn-in of 11,000 iterations. Autocorrelations, density plots, and history plots were examined for further evidence of convergence for the MixRM and Mix2PLM.

The Mix3PLM failed to converge after 35,000 iterations based on the Heidelberger and Welch diagnostics or on the ratio of the standard deviation of the parameter estimate to the MC standard error for the parameter estimate. Autocorrelation plots, density plots, and history plots also failed to show convergence. Based on the 95% credibility interval, however, the Mix3PLMs were considered to have converged after a burn-in of 6,000 iterations and a post burn-in of 11,000 iterations.

Model Selection. BIC was used to inform model selection. AIC was provided as a comparison index. The percentages in Table A1 of the Appendix indicate the number of correct model selection decisions for each condition in which the generating model was selected. Model selection for the MixRM was correct for all but one of the conditions for Constraint 1.

All model selections were 100 percent correct for the Mix2PLM. For the Mix3PLM, however, model selection results varied ranging between 25 percent and 100 percent correct. The lower percentages occurred mainly under Constraints 2 and 3. Results were similar under Constraint 1 for all conditions for all three MixIRTMs. Unlike results for the MixRM and Mix2PLM, however, there were clearly some problems for the Mix3PLM, with Constraint 2. For the smaller sample size, (N = 600), the Mix3PLM detected fewer correct 3-class models. The numbers of students for each class in the small sample (N = 600) with the 3-class condition were 360, 180, and 60, respectively. It is possible the smallest sample size among three classes (i.e., N = 60) might not have been sufficient to estimate the Mix3PLM. Further, increasing test length to 40 items but with the same smaller sample may not have provided sufficient additional information for accurate estimation of model parameters.

In addition, in the larger sample size (N = 2,400) and 1-class condition for the Mix3PLM with Constraint 2, only 45 percent correct detections were observed for both the 20- and 40-item tests. The 1-class solution is the usual IRT solution, i.e., with no latent classes. In this case, it appears that under Constraint 2, model selection did not work well for the usual 3PL model. These results suggest that Constraint 2 affected model selection when the larger sample size was simulated for the 1-class model (i.e., a 3PL model without any latent classes).

Comparison of model selections for AIC and BIC indicated BIC had more correct selections than AIC for the MixRM and Mix2PLM. This was also the case for the Mix3PLM except for three conditions: smaller sample size × 3-classes for Constraints 2 and 3 and the longer test length × larger sample size × 3-class using Constraint 3. These results are consistent with from Li et al. (2009) which found that BIC made more correct model selections for all three MixIRTMs.

Label Switching. Label switching was observed between replications. This type of label switching is easily observed in simulation studies because the generating parameters are known and can be compared with the estimated parameters for each of the latent classes (Cho, Cohen, & Kim, 2006; Li et al., 2009). When label switching was identified, the problem was solved by comparing frequencies between generated class membership and the posterior mode estimates of class membership. The latent classes of each replication were switched manually based on the frequency comparisons prior to the recovery analysis.

Recovery Analysis. Recovery was analyzed only for replications for which the correct number of latent classes was selected by BIC. For the MixRM, bias values were all zero for item difficulty and very small for latent class means, ranging between –.002 and .002. Correlations between generating values and item difficulty estimates were all high, ranging from .979 to .999. RMSEs for item difficulty ranged from .049 to .229.

For the Mix2PLM, the type of constraint did not appear to affect recovery of item difficulties (see Table A2 in the Appendix). RMSE results suggested recovery of the item and class mean parameters was also generally satisfactory. Bias and correlation statistics for item difficulty and discrimination estimates for the Mix2PLM were all zero for item difficulty. Recovery of item difficulty for the Mix2PLM was generally good with the possible exceptions of the small sample 3-class model conditions. Bias statistics for item discrimination parameters were also relatively small, ranging from .002 to .043. RMSEs in the large sample conditions were slightly smaller, ranging from .080 to .199 and suggest a sample size effect on recovery of item discrimination.

Bias and RMSE results for recovery analysis of class mean parameters in the Mix3PLM were all close to zero. Bias for the item difficulty and lower asymptote parameters also were close to zero as well, although bias values were higher than .3 for item discrimination in the 2,400 students × 1-class condition for Constraint 2 (person centering) and in the 40-items × 2,400 students × 1-class condition for Constraint 3 (item centering). In addition, Constraint 1 (item anchoring) had the lowest RMSE values for all conditions except for shorter test length × larger sample size × three classes for item difficulty and discrimination (see Tables A3 and A4 in the Appendix).

A high percentage of correct selections was observed for the MixRM and Mix2PLM with no apparent affect on model selection. The Mix3PLM, however, had a low percentage of correct selections under Constraints 2 and 3. These results appear to be related to poor recovery of item parameters.

Recovery of class membership was examined by calculating the percentage of correct class identifications for each condition and comparing that with the percentage of examinees simulated in that class after label switching. If the correct model was indicated by the BIC index, this was considered as a correct model selection. Latent class membership was recovered well for the MixRM and Mix2PLM. For the Mix3PLM, under Constraint 2, the 20 items × 2,400 students for the 2-class model had 84% correct identifications. All other conditions had correct identifications of 90% or greater. In addition, the percentage of correct membership identifications decreased as the number of latent classes increased. Sample size and type of constraint did not appear to affect class membership identification. These patterns were similar to those reported by Li et al. (2009).

Discussion and Conclusions

The effects of three scale identification methods were investigated for establishing a common metric between latent classes in MixIRT models. Results from an empirical example with the MixRM suggested that each of the constraints had a somewhat different effect on item difficulty estimates, ability estimates, numbers of latent classes, classifications of examinees into latent classes, and proportions of membership in each latent class. Similar results were observed for these data for the Mix2PLM with the exception that the same number of latent classes was extracted using all three constraints.

A simulation study investigated the impact of the three constraints in the context of three dichotomous MixIRT models: MixRM, Mix2PLM, and Mix3PLM. Exploratory MixRM, Mix2PLM, and Mix3PLM analyses were done to determine the best fitting model to the simulated data. The criterion used for model selection was BIC. Selection for the MixRM and Mix2PLM using BIC was close to 100 percent. For the Mix3PLM, model selection under Constraint 1 was better than under Constraints 2 or 3.

A recovery analysis was done to evaluate the effectiveness of the estimation algorithms for the different constraints. Item and latent class mean generating parameters were compared to the item and latent class mean estimates. Bias and RMSEs for latent class mean parameter estimates were close to zero for all conditions for all MixIRTMs. However, there were variations in recovery results for the different models.

For the MixRM, all bias and correlations suggested that generating parameters were recovered well. Type of constraint did not appear to affect recovery of item difficulties and recovery of item and class mean parameters was generally good. Recovery of item difficulties for the Mix2PLM was generally good except for the 3-class model in the small sample conditions. The type of constraint did not appear to affect recovery of item difficulty for this model. For the Mix3PLM, recovery was moderately good for Constraint 1 but less so for Constraints 2 and 3.

No constraint effect was observed on model selection for the MixRM or Mix2PLM. For the Mix3PLM, model selection under Constraint 1 was best compared to the other two constraints. Results suggest that any of the three constraints might be used for the MixRM and Mix2PLM but only Constraint 1 appeared appropriate for the Mix3PLM. Correct model selections and recovery were poorer for the Mix3PLM than for the other two MixIRT models. Latent class membership was recovered well for the MixRM and Mix2PLM.

Recovery was best for the Mix2PLM and worst for the Mix3PLM. When the types of constraints were compared, Constraint 2 had the worst results. Test length did not appear to affect recovery of item parameters although the longer test length was associated with improved correct identification of class membership in the MixRM and Mix2PLM. The larger sample size appeared to have better recovery of item parameters based on correlations. The more latent classes in the model, the poorer the recovery of class membership and item parameters.

Constraint 1 (Item Anchoring) generally performed best. An important problem, when using this constraint, is determining the set of anchor items. This is the same concern present in studies of differential item functioning (DIF) and equating. There are two methods to determine the anchor items in the MixIRTMs: deciding which items to select on the basis of theoretical reasons or based on statistical evidence (e.g., non-speededness items in speededness test). Choi et al. (2015) used the likelihood ratio test for DIF to determine the anchor items.

When three different types of constraint were involved in the present study, there was no problem for either the MixRM or the Mix2PLM. However, the Mix3PLM did not do as well with respect to selection of the correct model. This was particularly the case when Constraints 2 and 3 were used.

Acknowledgments

This paper is developed largely from the first author's unpublished dissertation (Choi, 2014), under the direction of the second author, and elements throughout are derived directly from that work.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the 2nd International Symposium on Information Theory* (pp. 267-281), Akademia Kiado, Budapest.

Baker, F. B. & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker. https://doi.org/10.1201/9781482276725

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*(4), 381-409. https://doi.org/10.3102/10769986026004381

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*(4), 331-348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with applications to DIF. *Journal of Educational and Behavioral Statistics, 35*(3), 336-370. https://doi.org/10.3102/1076998609353111

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2006, June). An investigation of priors on the probabilities of mixtures in the mixture Rasch model. Paper presented at the annual meeting of the Psychometric Society, Montreal, Canada.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2013). Markov Chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation, 83*(2), 278-306. https://doi.org/10.1080/00949655.2011.603090

Cho, S.-J., Cohen, A. S., & Templin, J. (2008, March). A multidimensional mixture IRT model for DIF analysis. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Choi, Y.-J. (2014). *Metric Identification in Mixture IRT Models*. (Unpublished doctoral dissertation). University of Georgia, Athens, GA.

Choi, Y.-J., Alexeev, N., Cohen, A. S., & Kim, S.-H. (2010, July). Identifying a Mixture 3-Parameter Model. Paper presented at the International Meeting of the Psychometric Society: The 75th annual meeting of the Psychometric Society, Athens, GA

Choi, Y.-J., Alexeev, N., & Cohen, A.S. (2015). DIF analysis using a mixture 3PL model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing, 15*(3), 239-253 https://doi.org/10.1080/15305058.2015.1007241

Congdon, P. (2003). *Applied Bayesian modelling*. New York: John Wiley. https://doi.org/10.1002/0470867159

Dai, Y., & Mislevy, R. (2006, April). Using structured mixture IRT models to study differentiating item functioning. Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, CA.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modelling of latent DIF. *Applied Psychological Measurement, 35*(8), 583-603. https://doi.org/10.1177/0146621611428446

Foy, P., Arora, A., & Stanco, G. M. (Eds.) (2013). *TIMSS 2011 user guide for the international database*. Chestnut Hill, MA: Boston College.

Heidelberger, P. & Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research, 31*(6), 1109-1144. https://doi.org/10.1287/opre.31.6.1109

International Association for the Evaluation of Educational Achievement (IEA). (2013). *TIMSS 2011 Assessment: Grade 8 Science Test.* [Data set]. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* Newbury Park, NY: Springer. https://doi.org/10.1007/978-1-4757-4310-4

Li, F., Cohen, A.S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement, 33*(5), 353-373. https://doi.org/10.1177/0146621608326422

Linacre, J. M. (2001a). *A user’s guide to WINSTEPS/MINISTEPS*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2001b). *Facets Rasch measurement software*. Chicago: Winsteps.com.

Linacre, J. M. & Wright, B. D. (2001). *A user's guide to BIGSTEPS*. Chicago, IL: MESA Press.

Mislevy, R., & Verhelst, N (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195-215. https://doi.org/10.1007/bf02295283

Muraki, E., & Bock, R. D. (2003). *PARSCALE* (Version 4.1) [Computer program]. Mooresville, IN: Scientific Software.

Muthén, L. K. & Muthén, B. O. (2012). *Mplus User's Guide*. Sixth Edition. Los Angeles, CA: Muthén & Muthén

Plummer, M., Best, N., Cowles, K., Vine, K., Sarkar, D., & Almond, R. (2012). *Package coda* [Computer software]. Available at http://cran.r-project.org/web/packages/coda/coda.pdf

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271-282. https://doi.org/10.1177/014662169001400305

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461-464. https://doi.org/10.1214/aos/1176344136

Sen, S., Cohen, A. S., & Kim, S.-H. (2016). The impact of non-normality on extraction of spurious latent classes in mixture IRT models. *Applied Psychological Measurement, 40*(2), 98-113. https://doi.org/10.1177/0146621615605080

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. J. (2007). *OpenBUGS User Manual*, Version 3.0.2., MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine, UK, available at http://mathstat.helsinki.fi/openbugs/Manuals/Contents.html.

Thissen, D., Chen, W-H., & Bock, R. D. (2003). *MULTILOG 7 for Windows: Multiple category item analysis and test scoring using item response theory* [Computer software]. Skokie, IL: Scientific Software International, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). DIF detection using IRT models. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

von Davier, M. (2001). *WINMIRA 2001 user's manual*. Available at http://208.76.84.140/ svfklumu/wmira/winmiramanual.pdf

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*(6), 389-406. https://doi.org/10.1177/0146621604268734

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton, NJ: Educational Testing Service.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG* (Version 3.0) [Computer program]. Mooresville, IN: Scientific Software.

Appendix A

Table A1. Percent of Correct Model Selections for the MixIRT Models

				BIC			AIC		
Constraint	Item	Sample	Latent Classes	Mix RM	Mix 2PLM	Mix 3PLM	Mix RM	Mix 2PLM	Mix 3PLM
1. Item Anchoring	20	600	1	100.00	100.00	100.00	95.00	100.00	100.00
			2	100.00	100.00	100.00	90.00	100.00	100.00
			3	100.00	100.00	100.00	95.00	100.00	100.00
		2400	1	100.00	100.00	100.00	**85.00**	100.00	100.00
			2	95.00	100.00	95.00	**70.00**	90.00	95.00
			3	100.00	100.00	95.00	**65.00**	95.00	95.00
	40	600	1	100.00	100.00	100.00	100.00	100.00	100.00
			2	100.00	100.00	100.00	100.00	100.00	100.00
			3	100.00	100.00	100.00	100.00	100.00	100.00
		2400	1	100.00	100.00	90.00	100.00	100.00	90.00
			2	100.00	100.00	100.00	100.00	100.00	100.00
			3	100.00	100.00	100.00	100.00	100.00	100.00
2. Person Centering	20	600	1	100.00	100.00	100.00	90.00	100.00	100.00
			2	100.00	100.00	100.00	90.00	100.00	100.00
			3	100.00	100.00	**40.00**	100.00	100.00	100.00
		2400	1	100.00	100.00	**45.00**	**65.00**	100.00	**45.00**
			2	100.00	100.00	90.00	**75.00**	95.00	**65.00**
			3	100.00	100.00	90.00	**55.00**	100.00	**60.00**
	40	600	1	100.00	100.00	90.00	100.00	100.00	**65.00**
			2	100.00	100.00	95.00	100.00	100.00	95.00
			3	100.00	100.00	**25.00**	100.00	100.00	100.00
		2400	1	100.00	100.00	**45.00**	100.00	100.00	**45.00**
			2	100.00	100.00	95.00	100.00	100.00	95.00
			3	100.00	100.00	95.00	100.00	100.00	**85.00**
3. Item Centering	20	600	1	100.00	100.00	100.00	95.00	100.00	100.00
			2	100.00	100.00	100.00	95.00	100.00	100.00
			3	100.00	100.00	**40.00**	100.00	100.00	100.00
		2400	1	100.00	100.00	100.00	**75.00**	100.00	100.00
			2	100.00	100.00	100.00	**75.00**	95.00	95.00
			3	100.00	100.00	100.00	**70.00**	95.00	90.00
	40	600	1	100.00	100.00	100.00	100.00	100.00	100.00
			2	100.00	100.00	95.00	100.00	100.00	95.00
			3	100.00	100.00	**25.00**	100.00	100.00	100.00
		2400	1	100.00	100.00	**25.00**	100.00	100.00	**25.00**
			2	100.00	100.00	**60.00**	100.00	100.00	**60.00**
			3	100.00	100.00	**30.00**	100.00	100.00	**70.00**

Table A1. (Cont'd).

				BIC			AIC		
Constraint	Item	Sample	Latent Classes	Mix RM	Mix 2PLM	Mix 3PLM	Mix RM	Mix 2PLM	Mix 3PLM
MixIRTM				98.86	100.00	82.36	91.25	99.17	88.06
Constraint 1				99.58	100.00	98.33	91.67	98.75	98.33
Constraint 2				100.00	100.00	75.83	89.58	99.58	79.58
Constraint 3				100.00	100.00	72.92	92.50	99.17	86.25
	20-items			99.72	100.00	88.61	82.50	98.33	91.39
	40-items			100.00	100.00	76.11	100.00	100.00	84.72
		N = 600		100.00	100.00	83.89	97.22	100.00	97.22
		N = 2400		99.72	100.00	80.83	85.28	98.33	78.61
			1-class	100.00	100.00	82.92	92.08	100.00	80.83
			2-class	99.58	100.00	94.17	91.25	98.33	91.67
			3-class	100.00	100.00	70.00	90.42	99.17	91.67

Table A2. Bias, RMSE and Correlations (Cor.) of Difficulty and Discrimination Parameters and Latent Group Mean in Mix2PLM over 20 Replications

				Difficulty			Discrimination			Latent Group Mean	
Constraint	Item	Sample	Latent Classes	Bias	RSME	Cor.	Bias	RSME	Cor.	Bias	RSME
1. Item Anchoring	20	600	1	0.000	0.104	0.996	0.039	0.157	0.947	0.003	0.015
			2	0.000	0.164	0.989	0.037	**0.228**	0.873	0.001	0.013
			3	0.000	**0.201**	0.984	0.014	**0.286**	0.800	0.005	0.024
		2400	1	0.000	0.060	0.999	0.034	0.092	0.985	-0.001	0.004
			2	0.000	0.097	0.996	0.042	0.143	0.955	-0.002	0.011
			3	0.000	0.123	0.994	0.043	0.181	0.922	0.000	0.009
	40	600	1	0.000	0.119	0.994	0.028	0.150	0.948	0.001	0.005
			2	0.000	0.162	0.990	0.023	**0.219**	0.882	0.001	0.007
			3	0.000	**0.211**	0.982	0.003	**0.265**	0.819	0.002	0.027
		2400	1	0.000	0.059	0.999	0.014	0.080	0.984	0.001	0.003
			2	0.000	0.092	0.997	0.015	0.118	0.965	-0.001	0.005
			3	0.000	0.128	0.993	0.025	0.171	0.928	0.000	0.009

Table A2. (Cont'd).

				Difficulty			Discrimination			Latent Group Mean	
Constraint	Item	Sample	Latent Classes	Bias	RSME	Cor.	Bias	RSME	Cor.	Bias	RSME
2. Person Centering	20	600	1	0.000	0.113	0.995	0.043	0.169	0.938	0.003	0.013
			2	0.000	0.172	0.988	0.037	**0.239**	0.862	0.001	0.005
			3	0.000	**0.222**	0.980	0.009	**0.313**	**0.776**	0.004	0.024
		2400	1	0.000	0.065	0.999	0.034	0.097	0.983	-0.001	0.004
			2	0.000	0.102	0.996	0.043	0.150	0.950	-0.002	0.018
			3	0.000	0.133	0.993	0.043	0.199	0.909	0.001	0.005
	40	600	1	0.000	0.123	0.994	0.028	0.156	0.943	0.001	0.004
			2	0.000	0.166	0.989	0.022	**0.224**	0.878	0.001	0.005
			3	0.000	**0.219**	0.981	0.003	**0.279**	0.807	0.001	0.008
		2400	1	0.000	0.062	0.999	0.013	0.083	0.983	0.001	0.002
			2	0.000	0.094	0.997	0.016	0.122	0.963	-0.001	0.011
			3	0.000	0.132	0.993	0.026	0.177	0.924	0.001	0.014
3. Item Centering	20	600	1	0.000	0.113	0.995	0.043	0.169	0.938	0.004	0.020
			2	0.000	0.172	0.988	0.038	**0.240**	0.861	0.000	0.018
			3	0.000	**0.221**	0.980	0.007	**0.316**	**0.772**	0.005	0.029
		2400	1	0.000	0.065	0.999	0.034	0.097	0.983	-0.004	0.019
			2	0.000	0.102	0.996	0.043	0.150	0.950	0.001	0.005
			3	0.000	0.134	0.993	0.043	0.199	0.908	0.001	0.011
	40	600	1	0.000	0.123	0.994	0.028	0.156	0.943	0.003	0.013
			2	0.000	0.166	0.989	0.023	**0.224**	0.878	0.001	0.008
			3	0.000	**0.219**	0.981	0.002	**0.280**	0.805	0.003	0.031
		2400	1	0.000	0.061	0.999	0.013	0.083	0.983	-0.001	0.003
			2	0.000	0.094	0.997	0.016	0.122	0.963	0.001	0.011
			3	0.000	0.132	0.993	0.026	0.177	0.924	0.001	0.012
Constraint 1				0.000	0.127	0.993	0.026	0.174	0.917	0.001	0.011
Constraint 2				0.000	0.134	0.992	0.026	0.184	0.910	0.001	0.009
Constraint 3				0.000	0.134	0.992	0.026	0.184	0.909	0.001	0.015
	20-items			0.000	0.131	0.992	0.035	0.190	0.906	0.001	0.014
	40-items			0.000	0.131	0.992	0.018	0.171	0.918	0.001	0.010
		N = 600		0.000	0.166	0.988	0.024	0.226	0.871	0.002	0.015
		N = 2400		0.000	0.096	0.996	0.029	0.136	0.953	0.000	0.009
			1-class	0.000	0.089	0.997	0.029	0.124	0.963	0.001	0.009
			2-class	0.000	0.132	0.993	0.030	0.182	0.915	0.000	0.010
			3-class	0.000	0.173	0.987	0.020	0.237	0.858	0.002	0.017

Note: When RMSE is larger than .2 or correlation is less than .8, the values are bold.

Table A3. Bias, RMSE and Correlations of Difficulty and Discrimination Item Parameters in Mix3PLM

				Difficulty			Discrimination		
Constraint	Item	Sample	Latent Classes	Bias	RSME	Cor.	Bias	RSME	Cor.
1. Item Anchoring	20	600	1	0.000	0.141	0.992	-0.032	0.196	0.917
			2	0.000	**0.215**	0.982	-0.089	**0.269**	0.829
			3	0.000	**0.290**	0.966	-0.174	**0.369**	**0.722**
		2400	1	0.000	0.099	0.996	-0.012	0.151	0.957
			2	0.000	0.168	0.989	-0.027	**0.208**	0.905
			3	0.000	**0.380**	0.942	-0.105	**0.295**	0.828
	40	600	1	0.000	0.173	0.988	-0.026	0.193	0.915
			2	0.000	**0.221**	0.981	-0.076	**0.260**	0.842
			3	0.000	**0.286**	0.967	-0.134	**0.323**	**0.775**
		2400	1	0.000	**0.480**	0.908	-0.148	**0.406**	0.842
			2	0.000	0.167	0.989	-0.050	0.189	0.920
			3	0.000	**0.256**	0.974	-0.064	**0.253**	0.859
2. Person Centering	20	600	1	0.000	0.162	0.990	-0.016	**0.212**	0.899
			2	0.000	**0.230**	0.979	-0.093	**0.279**	0.821
			3	0.000	**0.322**	0.958	-0.187	**0.383**	**0.720**
		2400	1	0.000	**0.828**	**0.724**	**-0.511**	**0.781**	**0.604**
			2	0.000	**0.439**	0.923	-0.096	**0.366**	0.829
			3	0.000	**0.349**	0.951	-0.127	**0.337**	**0.796**
	40	600	1	0.000	**0.335**	0.955	-0.134	**0.300**	0.883
			2	0.000	**0.324**	0.958	-0.108	**0.317**	0.814
			3	0.000	**0.294**	0.966	-0.138	**0.332**	**0.768**
		2400	1	0.000	**0.508**	0.897	**-0.309**	**0.450**	0.888
			2	0.000	**0.209**	0.983	-0.067	**0.245**	0.886
			3	0.000	**0.393**	0.939	-0.087	**0.282**	0.847
3. Item Centering	20	600	1	0.000	0.162	0.990	-0.013	**0.213**	0.897
			2	0.000	**0.227**	0.979	-0.092	**0.279**	0.821
			3	0.000	**0.304**	0.963	-0.174	**0.376**	**0.721**
		2400	1	0.000	0.118	0.995	-0.006	0.159	0.950
			2	0.000	0.161	0.990	-0.014	**0.216**	0.898
			3	0.000	0.192	0.985	-0.070	**0.263**	0.848
	40	600	1	0.000	0.183	0.987	-0.020	**0.202**	0.906
			2	0.000	**0.288**	0.967	-0.099	**0.298**	0.821
			3	0.000	**0.287**	0.967	-0.133	**0.330**	**0.768**
		2400	1	0.000	**0.584**	0.865	**-0.405**	**0.509**	0.876
			2	0.000	**0.497**	0.902	-0.228	**0.484**	0.824
			3	0.000	**0.354**	0.100	-0.053	**0.227**	0.097

Table A3. (Cont'd).

				Difficulty			Discrimination		
Constraint	**Item**	**Sample**	**Latent Classes**	**Bias**	**RSME**	**Cor.**	**Bias**	**RSME**	**Cor.**
Constraint 1				0.000	0.240	0.973	-0.078	0.259	0.859
Constraint 2				0.000	0.366	0.935	-0.156	0.357	0.813
Constraint 3				0.000	0.280	0.891	-0.109	0.296	0.786
	20-items			0.000	0.266	0.961	-0.102	0.297	0.831
	40-items			0.000	0.324	0.905	-0.127	0.311	0.807
		N = 600		0.000	0.247	0.974	-0.097	0.285	0.824
		N = 2400		0.000	0.343	0.892	-0.132	0.323	0.814
			1-class	0.000	0.314	0.941	-0.136	0.314	0.878
			2-class	0.000	0.262	0.969	-0.087	0.284	0.851
			3-class	0.000	0.309	0.890	-0.121	0.314	0.729

Table A4. Bias, RMSE and Correlations of Lower Asymptote Item Parameter and Latent Group Mean in Mix3PLM

				Lower Asymptote			Latent Group Mean	
Constraint	**Item**	**Sample**	**Latent Classes**	**Bias**	**RSME**	**Cor.**	**Bias**	**RSME**
1. Item Anchoring	20	600	1	0.025	0.068	**-0.177**	0.003	0.014
			2	0.027	0.075	**-0.421**	0.002	0.020
			3	0.028	0.078	**-0.489**	0.001	0.025
		2400	1	0.017	0.056	**0.342**	-0.001	0.006
			2	0.025	0.073	**-0.129**	-0.003	0.015
			3	0.021	0.078	**-0.317**	0.003	0.035
	40	600	1	0.025	0.068	**-0.057**	0.000	0.000
			2	0.029	0.074	**-0.308**	-0.001	0.013
			3	0.029	0.077	**-0.446**	0.005	0.038
		2400	1	0.020	0.059	**0.244**	0.000	0.001
			2	0.023	0.070	**-0.029**	-0.002	0.021
			3	0.027	0.079	**-0.207**	-0.002	0.023

Table A4. (Cont'd).

				Lower Asymptote			**Latent Group Mean**	
Constraint	**Item**	**Sample**	**Latent Classes**	**Bias**	**RSME**	**Cor.**	**Bias**	**RSME**
2. Person Centering	20	600	1	0.028	0.071	**-0.128**	0.003	0.011
			2	0.029	0.074	**-0.417**	0.001	0.008
			3	0.028	0.078	**-0.501**	0.003	0.019
		2400	1	0.075	**0.235**	**-0.347**	-0.001	0.004
			2	0.029	0.096	**-0.144**	-0.002	0.010
			3	0.027	0.087	**-0.280**	-0.005	0.038
	40	600	1	0.040	0.108	**-0.279**	-0.001	0.003
			2	0.034	0.090	**-0.319**	-0.002	0.013
			3	0.030	0.077	**-0.445**	0.005	0.051
		2400	1	0.046	0.153	**-0.266**	0.000	0.000
			2	0.027	0.077	**-0.042**	-0.001	0.008
			3	0.028	0.082	**-0.220**	0.000	0.021
3. Item Centering	20	600	1	0.029	0.071	**-0.131**	0.002	0.011
			2	0.028	0.074	**-0.429**	0.001	0.022
			3	0.030	0.077	**-0.496**	0.001	0.027
		2400	1	0.020	0.059	**0.308**	-0.001	0.004
			2	0.025	0.070	**-0.106**	-0.002	0.010
			3	0.026	0.074	**-0.302**	0.000	0.027
	40	600	1	0.028	0.070	**-0.052**	-0.001	0.003
			2	0.031	0.080	**-0.327**	-0.002	0.017
			3	0.030	0.077	**-0.450**	0.006	0.049
		2400	1	-0.002	0.130	**-0.388**	0.000	0.001
			2	0.062	0.167	**-0.223**	-0.004	0.019
			3	0.005	0.049	**-0.025**	0.002	0.039
Constraint 1				0.025	0.071	-0.166	0.000	0.018
Constraint 2				0.035	0.102	-0.282	0.000	0.016
Constraint 3				0.026	0.083	-0.218	0.000	0.019
	20-items			0.029	0.083	-0.231	0.000	0.017
	40-items			0.028	0.088	-0.213	0.000	0.018
		N = 600		0.029	0.077	-0.326	0.001	0.019
		N = 2400		0.028	0.094	-0.118	-0.001	0.016
			1-class	0.029	0.096	-0.078	0.000	0.005
			2-class	0.031	0.085	-0.241	-0.001	0.015
			3-class	0.026	0.076	-0.348	0.002	0.033