Original Research Article

Educational and Psychological Measurement
1–36

DOI: 10.1177/00131644221094325
journals.sagepub.com/home/epm
SAGE

The Impact of Sample Size and Various Other Factors on Estimation of Dichotomous Mixture IRT Models

Sedat Sen[1] and Allan S. Cohen[2]

Abstract

The purpose of this study was to examine the effects of different data conditions on item parameter recovery and classification accuracy of three dichotomous mixture item response theory (IRT) models: the Mix1PL, Mix2PL, and Mix3PL. Manipulated factors in the simulation included the sample size (11 different sample sizes from 100 to 5000), test length (10, 30, and 50), number of classes (2 and 3), the degree of latent class separation (normal/no separation, small, medium, and large), and class sizes (equal vs. nonequal). Effects were assessed using root mean square error (RMSE) and classification accuracy percentage computed between true parameters and estimated parameters. The results of this simulation study showed that more precise estimates of item parameters were obtained with larger sample sizes and longer test lengths. Recovery of item parameters decreased as the number of classes increased with the decrease in sample size. Recovery of classification accuracy for the conditions with two-class solutions was also better than that of three-class solutions. Results of both item parameter estimates and classification accuracy differed by model type. More complex models and models with larger class separations produced less accurate results. The effect of the mixture proportions also differentially affected RMSE and classification accuracy results. Groups of equal size produced more precise item parameter estimates, but the reverse was the case for classification accuracy results. Results suggested that dichotomous mixture IRT models required more than 2,000 examinees to be able to obtain stable results as even shorter tests required such large

[1]Harran University, Şanlıurfa, Turkey
[2]University of Georgia, Athens, GA

Corresponding Author:
Sedat Sen, Faculty of Education, Harran University, Osmanbey Kampusu, Sanliurfa 63300, Turkey.
Email: sedatsen06@gmail.com

sample sizes for more precise estimates. This number increased as the number of latent classes, the degree of separation, and model complexity increased.

Keywords

mixture item response theory, sample size, dichotomous data, maximum likelihood estimation, Monte Carlo simulation

Introduction

Since the seminal paper by Rost (1990), mixed Rasch models (MRMs; see also Kelderman & Macready, 1990; Mislevy & Verhelst, 1990) have attracted the attention of many researchers in the field of educational and psychological measurement (e.g., Bolt et al., 2001; Cohen & Bolt, 2005). Later, extensions of MRMs for both dichotomous and polytomous data have been developed and applied for several different testing situations (e.g., Austin et al., 2006; H. J. Cho et al., 2012; Cohen & Bolt, 2005; Sen, 2016). In essence, these models can be considered a mixture of item response theory (IRT) and latent class models. Combination of these two powerful statistical tools under one modeling approach provides many advantages. For example, qualitative information about the respondents can be obtained in addition to quantitative information about the items and respondents.

Mixture extensions of IRT models are mostly used for handling the heterogeneity behind the respondents (see Sen & Cohen, 2019 for a review). Traditional IRT models assume a single homogeneous population of respondents. However, it may not always be possible to collect data from homogeneous populations. Population of examinees may include two or more subpopulations that are captured in the different latent classes. Mixture IRT is known to be a useful modeling approach for the analysis of the heterogeneity in samples as it assumes that the overall population includes multiple latent classes that can be identified based on the item response patterns of respondents (Rost, 1990).

The mixture IRT modeling approach for dichotomous items has been used for several purposes in social science research (e.g., Alexeev et al., 2011; Bolt et al., 2002; Cohen et al., 2005; Maij-de Meij et al., 2010; Ölmez & Cohen, 2018; Sen, 2016). The parameters of the mixture IRT model include the parameters of each latent class and the IRT models. Thus, the final equation of mixture IRT resembles an IRT formulization with latent class properties. The probability of a correct response to an item under the three-parameter logistic mixture IRT (Mix3PL) model for dichotomous data can be given as follows:

$$\mathrm{P}\left(x_{ij}=1|\theta\right)=\sum_{g=1}^{G}\pi_{g}\left(\gamma_{ig}+\left(1-\gamma_{ig}\right)\frac{exp\left[\alpha_{ig}\left(\theta_{j}-\beta_{ig}\right)\right]}{1+exp\left[\alpha_{ig}\left(\theta_{j}-\beta_{ig}\right)\right]}\right), \tag{1}$$

where π_g represent the mixture proportions; θ_j denotes the latent ability for person j; and α_{ig}, β_{ig}, and γ_{ig} denote the discrimination, difficulty, and guessing parameters, respectively, for item i in class g. The two-parameter logistic mixture IRT (Mix2PL) model can be obtained from Equation 1 when the item guessing parameter is assumed to be zero. Similarly, the one-parameter logistic mixture IRT (Mix1PL) model can be obtained from Equation 1 by constraining of item discrimination estimates to be equal and guessing values to be zero. Due to their parsimony, the Mix1PL models (or the Rasch model variant, the mixture Rasch model, MRM) have been most frequently applied by researchers to explore the heterogeneous samples in diverse research contexts (Sen & Cohen, 2019). One of the advantages of these simpler models is that it is possible to obtain stable item difficulty parameter estimates with smaller sample sizes as the item discrimination and the pseudo guessing parameters are not estimated in these models. However, the equal discrimination assumption can be considered a disadvantage of the MRMs as it may lead to detection of spurious latent classes (see Alexeev et al., 2011). In some cases, Mix2PL and Mix3PL models may be more appropriate.

Mixture IRT models assume two or more unknown (latent) subclasses among the respondents. As these classes are not known a priori, an exploratory approach is appropriate in which the first step is to fit models with different numbers of classes to the same data and compare these models to select best fitting model. Relative fit indices (e.g., Akaike information criterion [AIC], Akaike, 1974; Bayesian information criterion, [BIC], Schwarz, 1978; the deviance information criterion [DIC], Spiegelhalter et al., 1998) can be used to determine the final model among the alternatives.

Model parsimony is a generally accepted principle in mixture IRT model selection, such that the simpler model is generally preferred over the more complex model. Model complexity also tends to increase with an increase in the numbers of latent classes and items. A recent review on mixture IRT applications by Sen and Cohen (2019) noted that the average sample sizes across 101 studies was 1,810.05 with a median of 1,000. The sample sizes of mixture IRT applications were found to range from $N = 99$ (Glück et al., 2002) to $N = 251{,}278$ (Oliveri & von Davier, 2011). For example, Glück et al. (2002) conducted an MRM application with 99 respondents. Such a small sample size, however, may not provide stable results as the MRM requires a high ratio of the number of persons to the test length (W. H. Finch & French, 2012). W. H. Finch and French (2012) suggest that for samples of 400 or fewer—mixture IRT models may not be particularly viable, except perhaps for the simplest models. Glück et al. (2002) also suggest that small sample size can be problematic for mixture IRT applications even with simpler models like the MRM.

As required sample size for stable results is an important factor to consider for statistical analyses, sample size must be considered at the study planning phase, because it impacts the accuracy and efficiency of model parameter estimates of IRT models (H. Finch & French, 2019). IRT models, in fact, are known to require larger sample size than other common statistical analyses such as t-test and ANOVA (Nye et al.,

2020). IRT models can produce stable results, however, with sample sizes as small as 100 (Cohen et al., 2001), although larger sample sizes would be needed for mixture extensions of IRT models.

Sample size is not typically a concern, when applying IRT models and their extensions in the context of large-scale assessments which by definition have large samples of examinees use country-based populations. It may not always be possible, however, for researchers to have large samples. For example, sample sizes are typically very small (<200) for samples of students with disabilities or children of migrant workers. In such cases, it could be difficult to accurately estimate the model parameters. More information is needed regarding methods for parameter estimation with small samples such as when working with these kinds of populations (H. Finch & French, 2019).

One concern with respect to sample size is that there is, as yet, little information reported on the use of mixture IRT models with smaller samples. Only a few simulation studies appear to have been reported within the mixture IRT framework on the effect of sample size. For example, a recent simulation study conducted by Kutscher et al. (2019) investigated the performance of mixture IRT models under different sample size conditions. The focus of their study was on sample sizes for two polytomous mixture IRT models, the restricted mixed generalized partial credit model (rmGPCM) and the mixed partial credit model (mPCM). The problems of estimation and accuracy of parameter and standard error estimates were examined by generating different sample sizes from 500 to 5,000 examinees in 500 step increments. Kutscher et al. (2019) suggested that the two mixture IRT models required at least 2500 observations for three latent class models to provide accurate parameter and standard error estimates. There are other simulation studies conducted on the performance of polytomous mixture IRT models under different data conditions (Y. Cho, 2014; Huang, 2016; Wetzel et al., 2016). For example, Huang (2016) generated mixed generalized partial credit model (GPCM) data sets with sample sizes of 200, 500, 1,000, and 2,000. Results for a mixture GPCM model showed optimum performance with 1,000 examinees and 20 items for a three-class solution. Similarly, Wetzel et al. (2016) conducted a simulation study with the mixed partial credit model (PCM), generating the data sets with different sample sizes including 200, 500, and 2,000. Results from Wetzel et al. indicated that all sample size conditions produced reasonable estimates for the one-class solution with 10-items and two-class solutions with five items. Y. Cho (2014) also examined the performance of a mixed PCM under sample size conditions of 1,200, 3,000, and 6,000. Results from Cho suggested that mixed PCMs with a four-class solution needed 3,000 cases and 10 items for stable results. Cho also reported that mixed PCMs with fewer classes required fewer than 3,000 cases. It should be noted that those studies specifically focused on polytomous versions of mixture IRT models.

Other simulation studies have also considered sample size when focusing on different aspects of dichotomous mixture IRT models. For example Li et al. (2009) investigated the performance of model selection indices for dichotomous mixture IRT models. Results from Li et al. reported that 600 examinees appeared to be

sufficient for one- to four-group MRMs and possibly for a Mix2PLM for 15- and 30-item tests with the BIC index. Frederickx et al. (2010) compared samples of 500 and 1,000 in a mixture IRT model with random items and reported the average misclassification rate was lowest in the larger sample size. W. H. Finch and French (2012) compared two estimation methods with sample sizes of 400, 1000, and 2,000. The researchers stated that both estimation methods had difficulty reaching convergence for many of the replications with the relatively small samples (e.g., 400). Results showed that more precise parameter estimates were obtained with larger samples and more items.

A range of sample sizes in simulation studies have been reported from 100 to 25,000 cases (S. J. Cho et al., 2010; W. H. Finch & French, 2012; W. Y. Lee et al., 2018; Maij-de Meij et al., 2010; Preinerstorfer & Formann, 2012). Sample sizes of 1,000, 1,500, and 2,000 were the most frequently used.

The effects of sample size on item parameter estimation has received attention in previous research (e.g., Baker, 1998; de la Torre & Hong, 2010; H. Finch & French, 2019; Swaminathan & Gifford, 1983; Swaminathan et al., 2003). There does not appear, however, to be much on dichotomous mixture IRT models even though, as noted above, mixture IRT models are known to fail with small sample sizes (e.g., W. H. Finch & French, 2012). Furthermore, results of simulation studies with polytomous mixture IRT models may not be appropriate as the type of model may have an effect on parameter recovery.

The objective of this study, therefore, is to examine the effects of sample sizes, and several other factors including test length, number of latent classes, the degree of latent class separation, and the mixture proportions for three dichotomous mixture IRT models, the Mix1PL, Mix2PL and Mix3PL, in the context of a comprehensive simulation study. The present study focuses specifically on examining the effects of these factors on estimation of item parameters and classification accuracy for dichotomous mixture IRT models.

Method

In this section, we report on a comprehensive simulation study designed to investigate effects of sample size, test length, number of latent classes, the degree of class separation, and the mixture proportions on the estimates of dichotomous mixture IRT model parameters and classification accuracy. The study used a fully crossed design consisting of 1,584 conditions: 11 sample sizes $\times$ 3 test lengths $\times$ 2 classes $\times$ 2 mixture proportions $\times$ 4 class separations $\times$ 3 dichotomous mixture IRT models. The sizes of each condition are reported in Table 1. The conditions used in this study were based on those reported in previous simulation studies in mixture IRT models (Alexeev et al., 2011; S. J. Cho et al., 2013; H. Finch & French, 2019; Jiao et al., 2012; Li et al., 2009; Maij-de Meij et al., 2010; Preinerstorfer & Formann, 2012).

Some research has suggested that a sample size of 100 or more may be adequate for estimation of model parameters for the 1PL IRT model (Wright, 1977). It would

Table 1. Manipulated Factors in Simulation Design.

Factor	Conditions
Sample size	100, 200, 300, 400, 500, 750, 1,000, 2,000, 3,000, 4,000, and 5,000
Test length	10, 30, and 50
# of class	2 and 3
Mixture Proportions	Equal (1/2, 1/2), nonequal (3/4, 1/4) for two-class solutionsEqual (1/3, 1/3, 1/3), nonequal (1/2, 1/4, 1/4) for three-class solutions
Degree of class separation	For 2-class solution: Large ability separation: $N(0, 1)$ and $N(1, 1)$; Medium ability separation: $N(0, 1)$ and $N(0.8, 1)$; Small ability separation: $N(0, 1)$ and $N(0.5, 1)$; Normal ability separation: $N(0, 1)$ and $N(0, 1)$For 2-class solution: Large ability separation: $N(0, 1)$, $N(1, 1)$, $N(-1, 1)$; Medium ability separation: $N(0, 1)$, $N(0.8, 1)$, $N(-0.8, 1)$; Small ability separation: $N(0, 1)$, $N(0.5, 1)$, $N(-0.5, 1)$; Normal ability separation: $N(0, 1)$, $N(0, 1)$, $N(0, 1)$
Model type	Mix1PL, Mix2PL, and Mix3PL IRT models

Note. IRT = item response theory.

be useful to study whether this would be sufficient for mixture 1PL models. Accordingly, we chose a set of sample sizes included 100, 200, 300, 400, 500, 750, 1,000, 2,000, 3,000, 4,000, and 5,000 respondents. Sample size values less than 500 were selected to examine the effect of relatively small sample sizes on parameter estimates for the mixture models described in Table 1. Sample sizes larger than 500 were selected as these have been reported in previous simulation studies (e.g., S. Lee et al., 2021; Li et al., 2009). We also generated three different test lengths: 10, 30, and 50 items, as reported in S. J. Cho et al. (2013). These test lengths have also been used in a number of simulation studies (e.g., H. Finch & French, 2019; Li et al., 2009). These three test lengths simulated tests having a small, medium, or large number of items. Two different numbers of latent classes (two and three latent) were generated. Two mixture proportions were simulated: equal and nonequal proportions in each latent class. Equal proportions were $\pi_1 = \pi_2 = 1/2$ for two-class models and $\pi_1 = \pi_2 = \pi_3 = 1/3$ for three-class models as reported in S. J. Cho et al. (2013). The ratio for two-class solutions of 0.25–.75 (Maij-de Meij et al., 2010) was used for two-class models and 0.25-0.50-0.25 (Cassiday et al., 2021) was used for three-class nonequal mixture proportion models. Thus, for the equal proportion two-class conditions, each class was simulated to consist of 50% of the examinees (e.g., if 1,000 examinees was simulated, each latent class was constrained to consist of 500 examinees). Four different degrees of latent class separation (large, medium, small and normal) were generated by manipulating the ability distributions of each class following the method described in Jiao et al. (2012). For the large class separation conditions in two-class solutions, ability parameters for Class 1 were simulated from a standard

Table 2. Generating Item Parameter Values for 10-Item Conditions.

Item	α_1	α_2	α_3	β_1	β_2	β_3	γ_1	γ_2	γ_3
1	1	2	1	–0.436	1.443	0.482	0.10	0.25	0.20
2	1	2	1	–0.694	1.168	1.138	0.10	0.25	0.20
3	1	2	1	–0.848	0.761	0.038	0.10	0.25	0.20
4	1	2	2	–1.040	0.900	–1.782	0.20	0.20	0.25
5	1	2	2	–1.646	–1.249	–0.978	0.20	0.20	0.25
6	2	1	2	1.215	–0.120	–0.294	0.20	0.20	0.25
7	2	1	2	1.145	–0.195	–0.408	0.20	0.20	0.25
8	2	1	1	1.375	–0.036	1.419	0.25	0.10	0.10
9	2	1	1	0.936	–0.552	0.593	0.25	0.10	0.10
10	2	1	1	0.333	–1.304	0.865	0.25	0.10	0.10

Note. α =item discrimination; β = item difficulty; γ = item guessing.

normal distribution $N(0, 1)$, whereas those for Class 2 were simulated from $N(1, 1)$. While the distribution remains the same for Class 1, the ability parameters of Class 2 were simulated from $N(0.8, 1)$ and $N(0.5, 1)$ for medium and small separation conditions, respectively. However, two classes were generated to have the same distribution $N(0, 1)$ for the normal separation (i.e., no separation) condition. A similar approach was used for three-class conditions (see Table 1). Three different dichotomous mixture IRT models were considered: Mix1PL, Mix2PL, and Mix3PL models.

Data were generated for each of the three types of dichotomous mixture IRT models using code written in R, version 4.0.3 (R Core Team, 2021). The distributions of examinee ability and item parameters (i.e., discrimination, difficulty, and guessing) were generated to be the same for each of the three models. Person ability parameters were randomly drawn from a standard normal distribution with a mean of 0 and a standard deviation of 1 as in Alexeev et al. (2011). However, class-specific item parameters were generated for each model and item parameter values for the classes were made to be different to differentiate the classes in the simulations (see Table 2). Item difficulty values were randomly drawn from a standard normal distribution $N(0,1)$ between -2 and 2 as in DeMars and Lau (2011) and Preinerstorfer and Forman (2012). Generating values of item discrimination parameter included two values: 1 (for poor performance items) and 2 (for good performance items) as in Li et al. (2009). Values of item guessing parameters were .10, .20, and .25 as in Li et al. (2009). These values represent easy items, medium difficulty items, and difficult items, respectively. Table 2 shows the generating parameters for the 10 item conditions. (Generating item parameter values for 30- and 50-item conditions are presented in the Appendix.) The values provided in Table 2 were used to generate two- and three-class solutions for each of the three different mixture IRT models. For example, only item difficulty values of first two columns (–0.436 and 1.443 were used for Item 1, so on so forth) were used for generating values for two-class Mix1PL conditions. Similarly item difficulty values of three columns (–0.436, 1.443, and 0.482 were used

for Item 1, so on so forth) were used for generating values for three-class Mix1PL conditions. Different model and class conditions can be obtained from this table with the same logic. One hundred replications were simulated for each condition and a total of 158,400 data sets were generated.

After the data were simulated, all data sets were analyzed for each of the mixture IRT models with the computer program Mplus version 8.6 (Muthén & Muthén, 1998-2021) using the robust maximum likelihood estimation (REML). This version of Mplus was selected because it can be used to estimate all of the dichotomous mixture IRT models considered in this study.

Convergence problems are sometimes observed in the ML estimation of mixture IRT models, particularly with the more complex conditions (e.g., for the three-class Mix3PL model). For example, local maxima may be found in the global ML solution. For this reason, it might be useful to use either the true (i.e., generating) parameter values or multiple random values as the starting values, although this does not guarantee convergence. In this study, the generating item parameters were used as the starting values. For example, instances of solutions with extreme item parameter estimates were observed for some conditions. The outputs with extreme item parameters beyond the expected scale for the parameter (e.g., 1,903.84 for item difficulty) were excluded from and new data were generated.

To evaluate the accuracy of item parameter estimates, it is necessary to compare the estimated values with the ''true'' (generating) values. In this study, recovery of item parameter estimates was assessed using root mean square errors (RMSEs) calculated by taking the square root of the mean of squared deviations of estimated parameter values minus their generating values. For example, the RMSE for the item difficulty parameter for a specific condition was calculated as,

$$RMSE_{(\beta_i)} = \sqrt{\frac{\sum_{r=1}^{R} \left(\hat{\beta}_i - \beta_i\right)^2}{R}}, \tag{2}$$

where R ($r = 1, \ldots R$) is the number of replications, β_i denotes true parameter value, and $\hat{\beta}_i$ denotes the estimated parameter value for item i. Equation 2 was also used for assessment of item discrimination and item guessing parameter estimates. Before calculating the RMSE for a given replication, parameter estimates were first transformed to the scale of the generating values. The parameter estimates are exactly the same as the true value when RMSE equals zero. Lower values (e.g., <0.10) indicate better fit. In addition to RMSE, the bias and the mean absolute error (MAE) values were also computed using *metrics* package (Hamner et al., 2018) in R. These are not reported in the main body of the paper due to space limitations, but are reported in full in the supplementary material.

Classification accuracy was also evaluated. In addition to item parameters, estimation of mixture IRT models yields a posterior probability for each person in each latent class based on his or her response pattern (Maij-de Meij et al., 2008). The person is classed in the latent class for which the posterior probability is the highest. For

example, if two posterior probability values of a specific person were 0.91 and 0.09 for Class 1 and Class 2, respectively, this person would be assigned to Class 1. These values are saved in the Mplus output using the SAVEDATA command along with the FILE option (e.g., SAVEDATA: FILE = ''respondents1.cprob'';). Simulated latent class membership values were compared with estimated highest probability latent class membership for each replication. A percentage of correct membership was calculated by determining the number of matched assignments between these two set of values. Estimated latent class membership values were saved in the Mplus output and extracted with the *MplusAutomation* package (Hallquist & Wiley, 2018). For a data set with 1,000 examinees, classification accuracy value was calculated as 0.95 or 95 percent, if there was a matched assignment for 950 of the 1,000 cases.

It should be noted that class labels of different data sets may change during the simulation studies as the designation of classes in Mplus arbitrary. This is referred to as label switching. For example, given that Class 1 was simulated, the estimated latent class might be Class 2. To avoid this problem, we used starting values for each item parameter in Mplus syntax. In this case, label switching would also occur. To determine whether label switching occurred, we first compared item parameter estimates for each latent class with the generating values for a given replication. We re-labeled the estimated latent classes, so that the estimated item parameters were closest to the generating values for each latent class. We did not change item parameter estimates in making this correction. In this way, we used the item parameter estimates to determine whether label switching had occurred. Thus, if label switching had occurred such that Class 1 was labeled Class 2 in an output file, the classes were re-labeled before tabulating RMSE and classification accuracy results. This was done for each replication for each number of latent classes condition. RMSE values and classification accuracy percentages were both calculated once any label switching was corrected.

Results

Results of recovery of generating values for item parameters and classification accuracy of mixture proportions were obtained for the 1,584 conditions in the study. Mean RMSE values and percentage of classification accuracy values were computed for the 100 replications for each condition. Figures 1 to 6 summarize the mean RMSE results for each fitted model. Separate plots are provided for item discrimination, difficulty and guessing parameters in each figure. Figures 7 to 9 summarize the percentage of classification accuracy results for each fitted model. Each figure includes eight plots, four of which are for two-class solutions and four others which are for three-class solutions, with four different degrees of separation: (a) large, (b) medium, (c) small, and (d) normal. Plots in each figure also display six labeled lines representing six different conditions for the three test lengths (10, 30, and 50 items) and two mixture proportions (E stands for equal proportions, NE stands for nonequal proportions). For example, E50 indicates a condition with equal group size in the 50-

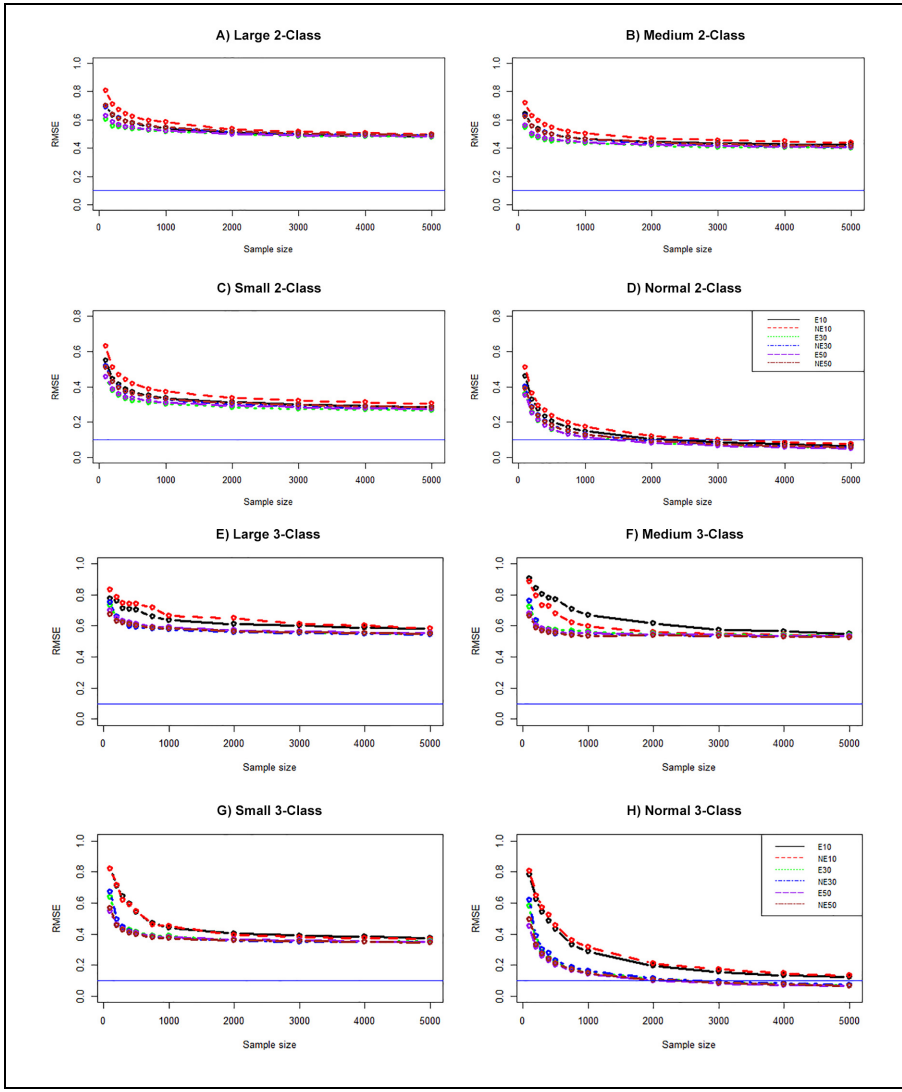

Figure 1. Mean RMSE Values for Item Difficulty Parameter Under the Mix1PL Model
Note. RMSE = root mean square error.

item condition. Eleven sample size conditions are specified along the x-axis of the plots.

Item Parameter Recovery Results

Minimum and maximum RMSE values of item difficulty parameter estimates for the two-class solutions are presented in Table 3A (see Appendix). As shown in

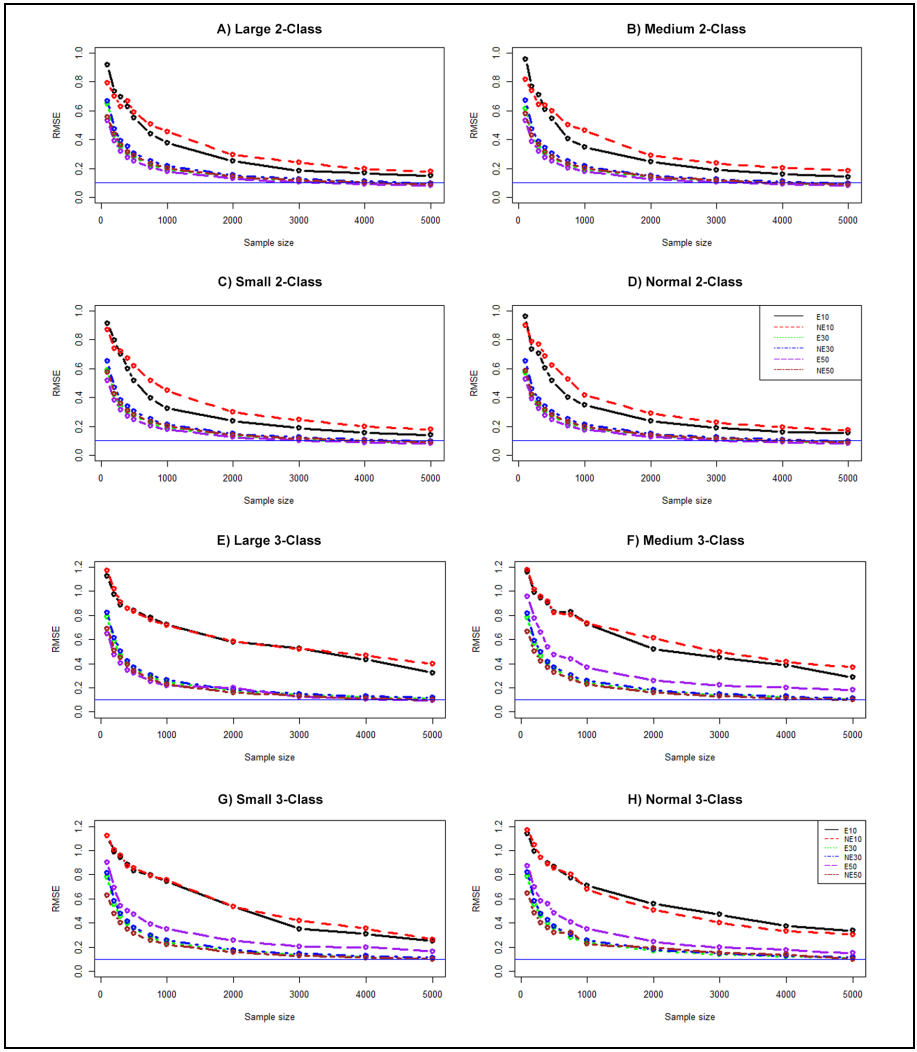

Figure 2. Mean RMSE Values for Item Discrimination Parameter Under the Mix2PL Model.
Note. RMSE = root mean square error.

Table 3A, mean RMSE values for the item difficulty parameter ranged from 0.051 to 0.511 for the normal class separation conditions with two-class solutions. The ranges of the mean RMSE values for the item difficulty parameter were computed for small (0.265–0.629), medium (0.399–0.720), and large (0.478–0.806) class separation conditions for the two-class solutions. Figure 1 displays the mean RMSE values for item difficulty parameter for the Mix1PL model for each condition. As expected, average

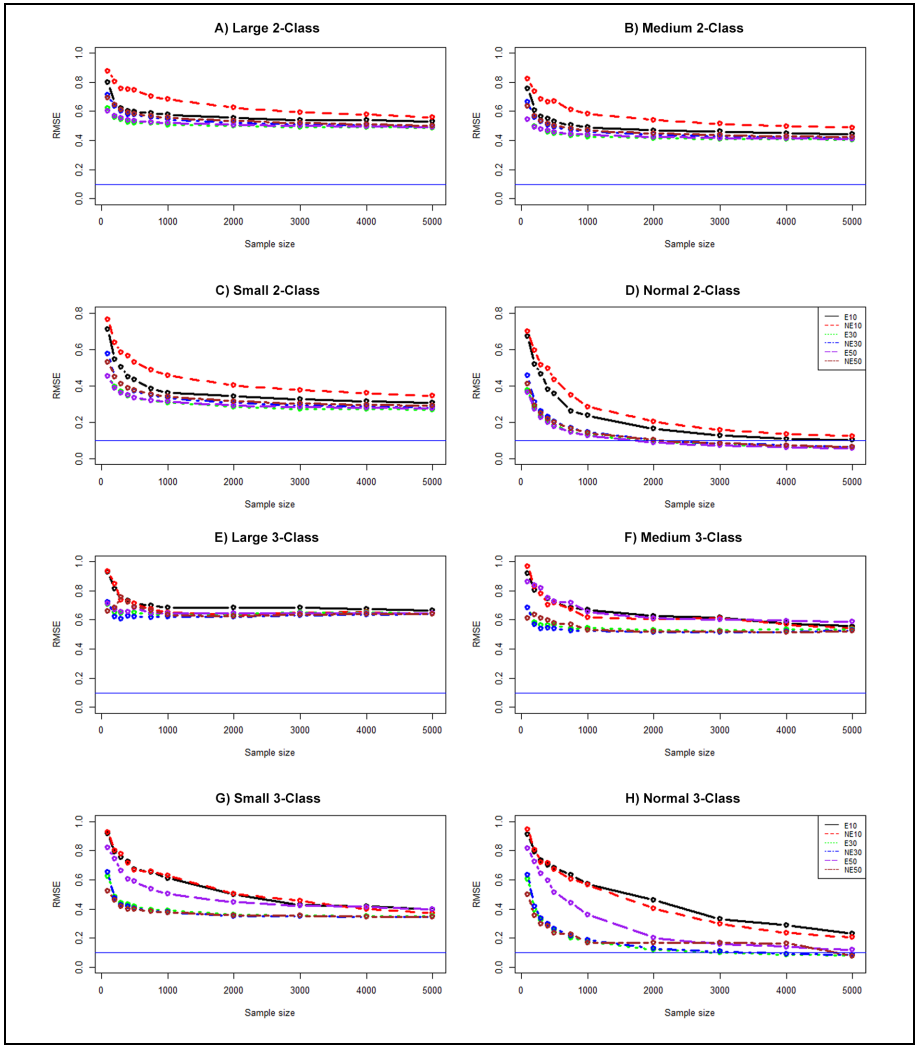

Figure 3. Mean RMSE Values for Item Difficulty Parameter Under the Mix2PL Model.
Note. RMSE = root mean square error.

RMSE values decreased as the sample size increased in two- and three-class conditions. As shown in Figure 1A to 1D, 10-item conditions with nonequal group sizes (NE10) produced the highest mean RMSE values and 30- and 50-item conditions with nonequal group sizes (NE30 and NE50) produced the lowest mean RMSE values for normal latent class separation conditions. Similar patterns were observed with the conditions with other class separations in terms of the highest and lowest RMSE

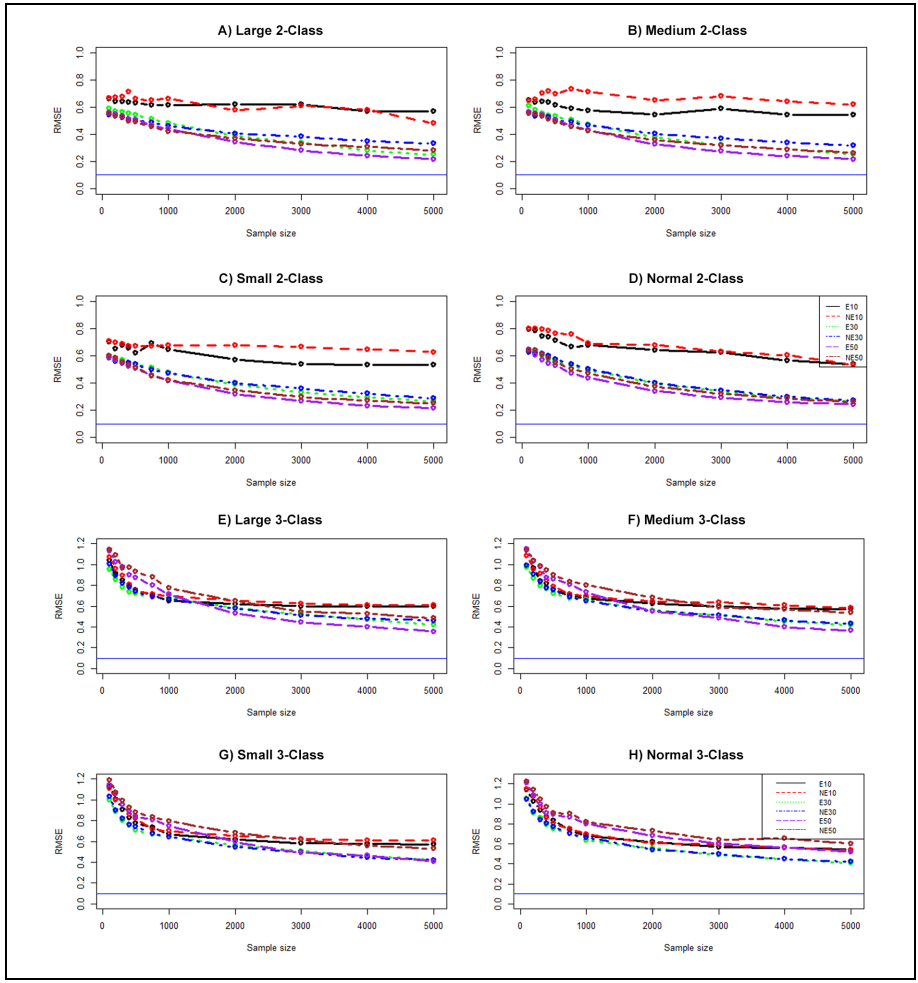

Figure 4. Mean RMSE Values for Item Discrimination Parameter Under the Mix3PL Model.

means. Recovery of item difficulty parameters across all conditions appeared to be affected by the test length. Increase in the test length appears to have a positive effect on the recovery of item difficulty parameters. As can be seen, 30- and 50-item conditions had mean RMSE values that were lower than the 10-item conditions. For the mixture proportions, the recovery of item difficulty parameters for equal proportions was better than that of nonequal mixture proportions. As shown in Figure 1A to 1D, only the normal class separation conditions had mean RMSE values less than 0.10 for 2,000 and more examinees (see the horizontal line). Mean RMSE values of all of the conditions were below 0.10 when the sample size was 3,000 examinees and were

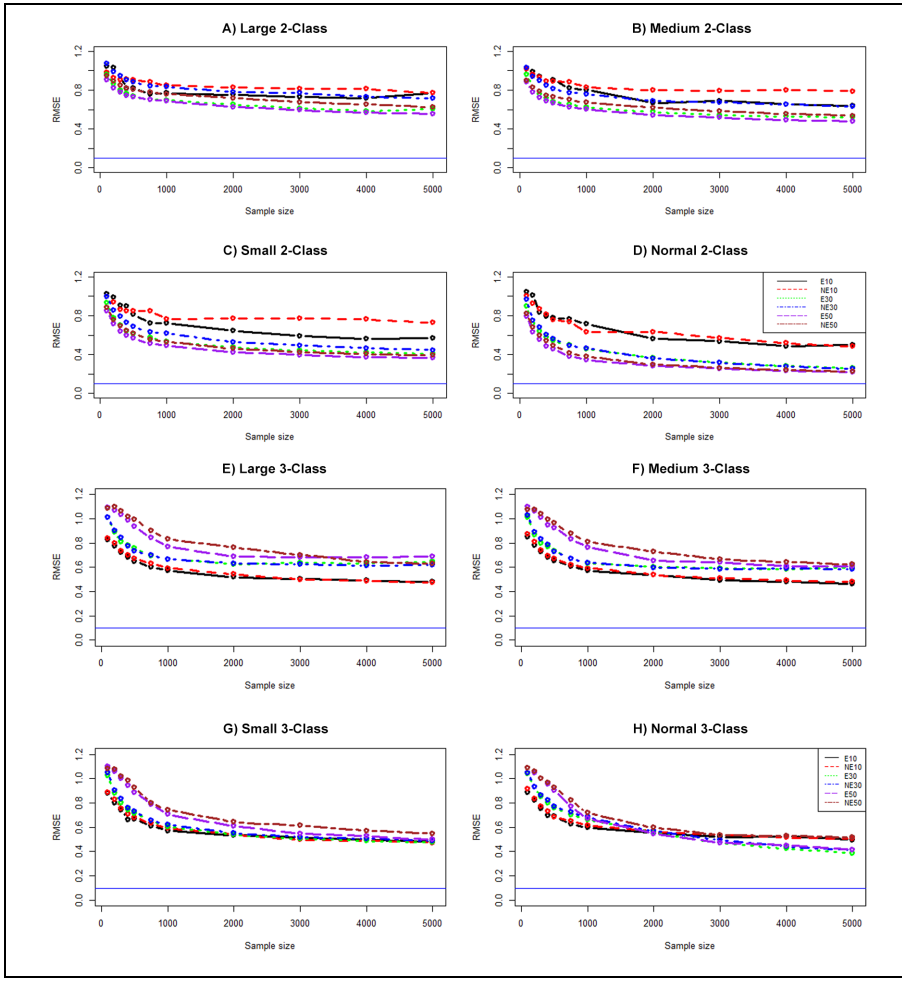

Figure 5. Mean RMSE Values for Item Difficulty Parameter Under the Mix3PL Model.
Note. RMSE = root mean square error.

large for conditions with less than 500. Similar patterns were observed for the conditions with other class separations with respect to the sample size effect. However, none of these class separation conditions produced mean RMSE values lower than 0.10, even with the largest sample size conditions (i.e., 5000). The mean RMSE values appeared to increase as the degree of class separation increased from small to large.

Minimum and maximum RMSE values of item difficulty parameter estimates for the three-class solutions are also presented in Table 3A (see Appendix). As shown in Table 3A, mean RMSE values for the item difficulty parameter ranged from 0.064 to

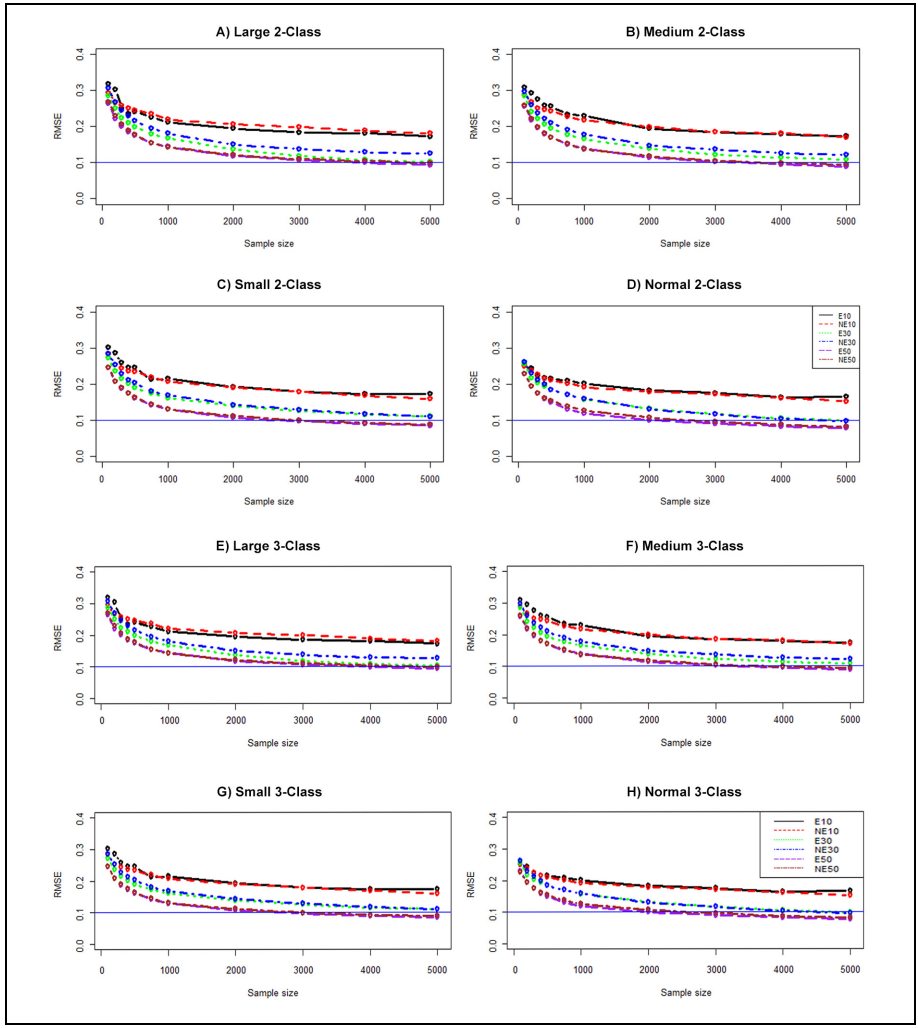

Figure 6. Mean RMSE Values for Item Guessing Parameter Under the Mix3PL Model.
Note. RMSE = root mean square error.

0.806 for the normal class separation conditions with three-class solutions. The ranges of the mean RMSE values for the item difficulty parameter were higher for small (0.345–0.823), medium (0.528–0.905), and large (0.546–0.834) class separation conditions for the three-class solutions. As shown in the plots in Figure 1E to 1H, the 10-item conditions with equal- and nonequal group sizes (E10 and NE10) produced the highest mean RMSE values and the 50-item conditions with both equal and nonequal group sizes (E50 and NE50) produced the lowest mean RMSE values.

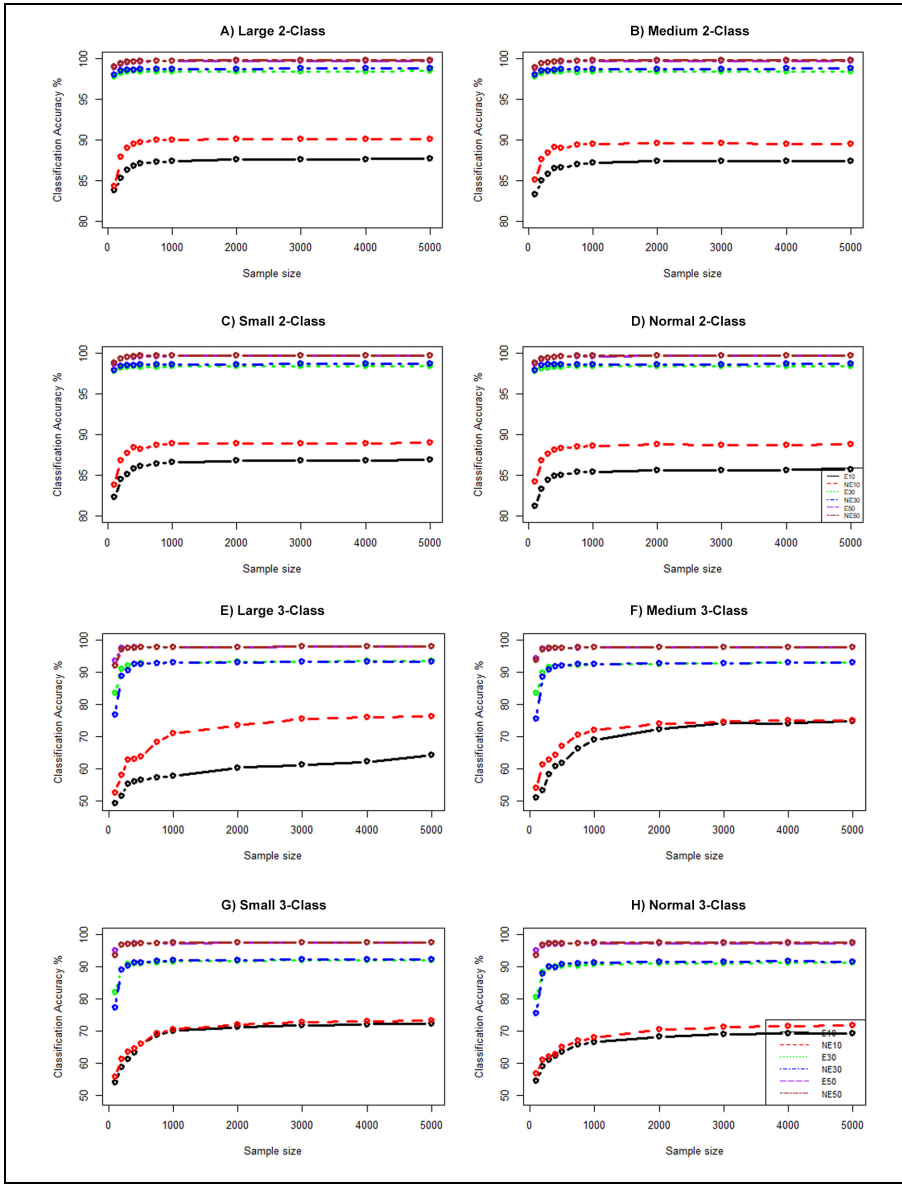

Figure 7. Classification Accuracy Percentages Under the Mix1PL Model.

Recovery of item difficulty parameters across all conditions appeared to be affected by the test length. Increase in test length had a positive effect on recovery of item difficulty parameter estimates. Mixture proportions, however, appeared to differ

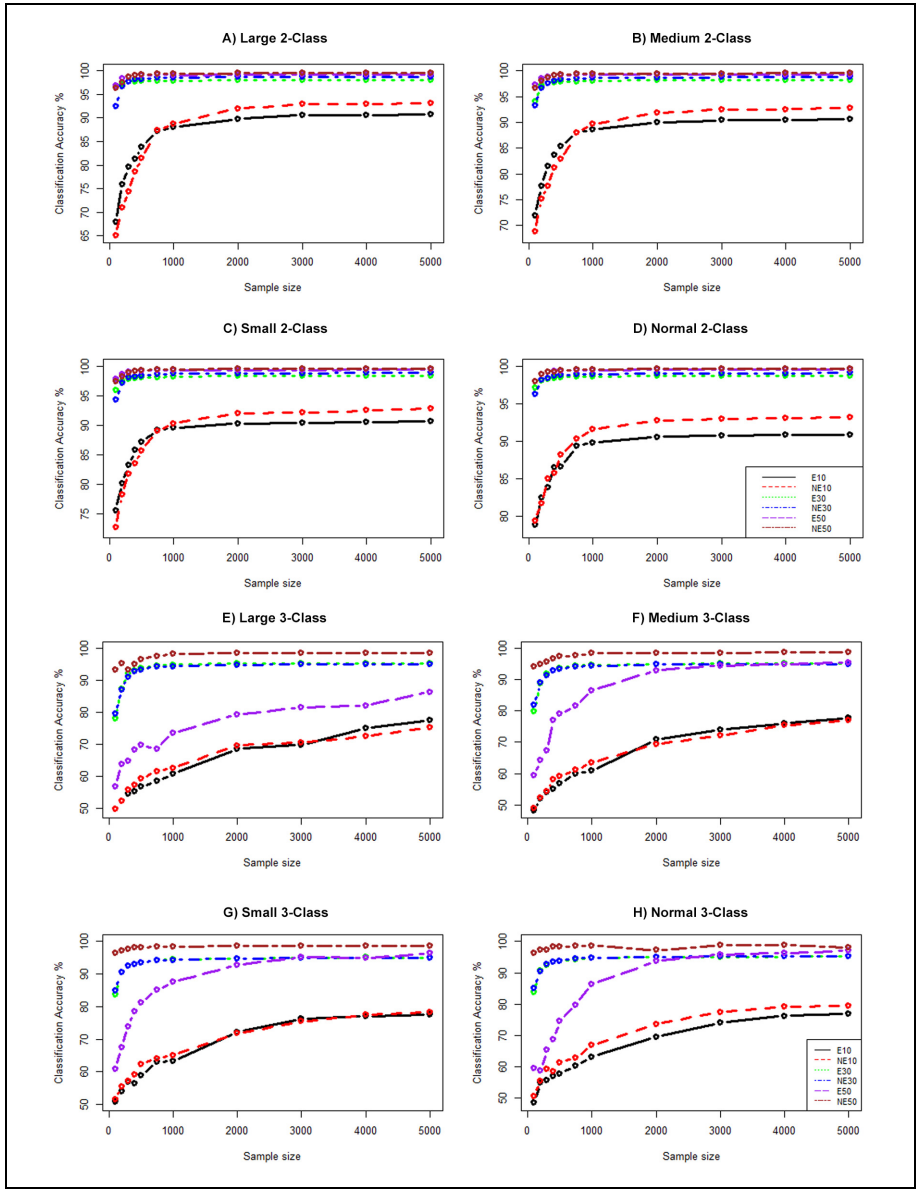

Figure 8. Classification Accuracy Percentages Under the Mix2PL Model.

depending on the recovery of item difficulty parameters. For example, nonequal group sizes had higher RMSE values for 10-item normal and large class separation conditions but the reverse for the 50-item conditions.

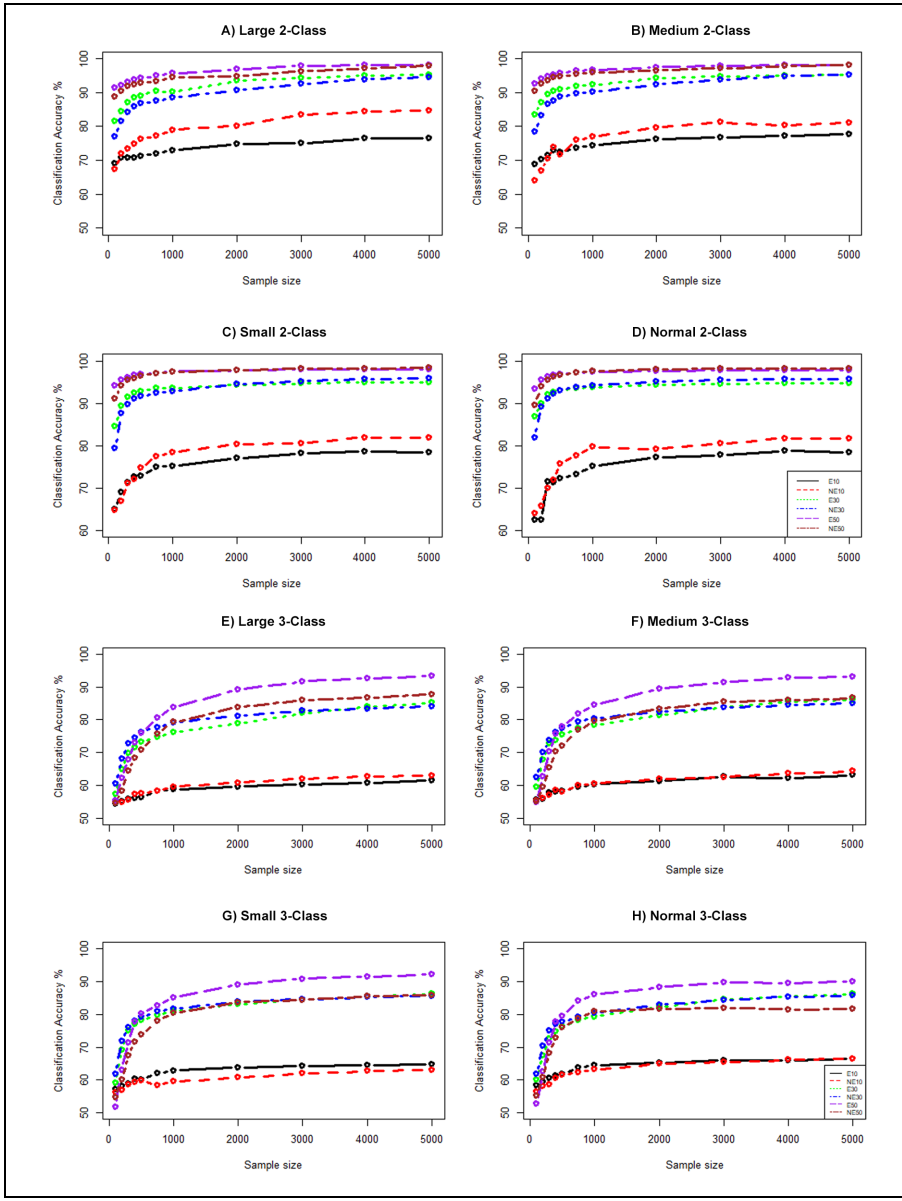

Figure 9. Classification Accuracy Percentages Under the Mix3PL Model.

Recovery of item difficulty parameters across all conditions appeared to be affected by the degree of class separation. Increase in the degree of class separation from small to large had a negative effect on recovery of item difficulty parameter

estimates. As shown in the plots in Figure 1E to1H, none of the conditions yielded mean RMSE values less than 0.10 with small, medium and large class separations (see horizontal line). Mean RMSE values for all conditions with 30 and 50 items were below 0.10, when the sample size was 3,000 or greater for the normal class separation conditions. Similar to the two-class solutions, mean RMSE values were larger for the conditions with fewer than 500 examinees and three classes. Overall, recovery in conditions with the three-class Mix1PL were worse, than results of the two-class Mix1PL. The three-class conditions appeared to have the worst recovery, particularly in the small sample with large class separation, 10-item conditions. The number of latent classes simulated, sample size, test length, the degree of class separation, and to some extent, class size appeared to affect recovery of item difficulty parameters for the Mix1PL model.

Minimum and maximum RMSE values of item discrimination parameter estimates for the two-class solutions of Mix2PL model are presented in Table 4A (see Appendix). As shown in Table 4A, mean RMSE values for the item discrimination parameter ranged from 0.081 to 0.961 for the two-class solutions with normal class separation. The ranges of the mean RMSE values for the item discrimination parameter were computed to be lower for small (0.080–0.915), medium (0.081–0.959), and large (0.081–0.918) class separation conditions for the two-class Mix2PL model solutions. The plots in Figure 2A to 2D display the mean RMSE values for item discrimination parameter under the Mix2PL model for each condition. Average RMSE values decreased as the sample size increased in both two- and three-class conditions. As shown in Figure 2A to 2D, the 10-item conditions with both equal- and nonequal group sizes (E10 and NE10) produced the highest mean RMSE values and 50-item conditions with equal and nonequal group sizes (E50 and NE50) yielded the lowest mean RMSE values for item discrimination parameters. Recovery of item discrimination parameters across all conditions appeared to be affected by the test length. Increase in the test length also had a positive effect on the recovery of item discrimination parameter estimates. Conditions with 30 and 50 items showed better recovery in terms of item discrimination. Overall, nonequal latent group size conditions had higher RMSEs than conditions with equal group sizes for each test length. As shown in Figure 2A to 2D, none of the six conditions yielded mean RMSE values less than 0.10 with 3,000 or fewer examinees (see horizontal line). Mean RMSE values of some of the 30 and 50 item conditions were below 0.10 when sample size was around or above 4,000. Mean RMSE values were large for conditions with fewer than 1,000 examinees, particularly with short tests.

The degree of latent class separation, however, appeared to differentially affect the recovery of item discrimination parameters. For example, recovery in the normal separation conditions yielded smaller RMSE values for the large sample size conditions (i.e., 1,000 or more) but the reverse was the case for the conditions with fewer than 1,000 examinees. The mean RMSE values for item discrimination parameters were close to each other across the different class separation conditions, such that. No clear pattern was observed.

Minimum and maximum RMSE values of item discrimination parameter estimates for the three-class solutions are presented in Table 4A (see Appendix). As shown in Table 4A, mean RMSE values for the item discrimination parameter ranged from 0.099 to 1.169 for the normal class separation with three-class solutions. The mean RMSE values for the item discrimination parameter ranged from 0.101 to 1.121 for the small separation condition. The range for the three-class Mix2PL model solutions was 0.102 to 1.175 for the medium separation condition, and 0.097 to 1.171 for the large class separation conditions. As shown in Figure 2E to 2H, 10-item conditions with nonequal group sizes (NE10) produced the highest mean RMSE values and 50-item conditions with nonequal group sizes (NE50) yielded the lowest mean RMSE values. Recovery of item discrimination parameters across all conditions appeared to be affected by the test length. Increase in the test length had a positive effect on the recovery of item discrimination parameter estimates. Mixture proportions, however, appeared to differ depending on the recovery of item discrimination parameters. As shown in the Figure 2E to 2H, none of the three-group conditions yielded mean RMSE values less than 0.10 with 4,000 or fewer examinees (see horizontal line). Mean RMSE values of some of the conditions with 50 items were below 0.10 when the sample size was around 5,000. Similar to the two-class solutions, mean RMSE values were large for the conditions with less than 1,000 examinees in the three-class conditions. Overall, conditions for the three-class Mix2PL model had worse recovery, compared with the results of the two-class Mix2PL model. The three-class conditions had the worst recovery, particularly in the small sample, 10-item test condition. To briefly summarize, the number of latent classes, sample size, test length and to some extent, the degree of class separation and class size appeared to affect recovery of item discrimination parameters of the Mix2PL models.

The results for item difficulty parameters of the Mix2PL model are summarized in Figure 3. RMSEs were similar to those in Figure 2. As expected, average RMSE values for item difficulty parameter decreased as the sample size and test length increased in two- and three-class conditions.

Minimum and maximum RMSE values of item difficulty parameter estimates for the two- and three-class solutions are presented in Table 5A (see Appendix). As shown in Table 5A, mean RMSE values for the item difficulty parameter ranged from 0.057 to 0.698 for the two-class solutions with normal class separation. The mean RMSE values for the item difficulty parameter ranged from 0.269 to 0.762 for the small separation condition, 0.403 to 0.822 for the medium separation condition, and 0.485 to 0.878 for the large class separation conditions for the two-class Mix2PL model solutions. The mean RMSE values for the item difficulty parameter ranged from 0.076 to 0.947 for the three-class solutions for the normal class separation conditions. The mean RMSE values for the item difficulty parameter ranged from 0.343 to 0.928 for the small separation condition, 0.514 to 0.969 for the medium separation condition, and 0.607 to 0.934 for the large class separation condition for the three-class Mix2PL model solutions. As shown in Figure 3, the three-class Mix2PL model conditions had worse recovery compared with the results of those for the two-class

Mix2PL model. RMSEs were lower for the 30- and 50-item conditions with equal group sizes (E30 and E50) and higher for the 10-item both equal- and nonequal group conditions (E10 and NE10) for both the two- and three-class solutions. RMSEs decreased as test length increased for recovery of item difficulty parameter estimates. Overall, nonequal group sizes had higher RMSEs than equal group sizes for each test length. As shown in Figure 3A to 3D, no mean RMSE values were lower than 0.10 for the small, medium, or large class separation conditions. Mean RMSE values of some of the conditions with 30 and 50 items, however, were lower than 0.10, for the sample size of 3,000 or greater for both two- and three-class model solutions for the normal class separation conditions. Mean RMSE values were large for samples with fewer than 3,000 examinees, particularly for the short tests.

Minimum and maximum RMSE values of item discrimination parameter estimates for the two- and three-class Mix3PL model solutions are presented in Table 6A (see Appendix). As shown in Table 6A, mean RMSE values for the item discrimination parameter ranged from 0.244 to 0.801 for the two-class solutions with normal class separation. The mean RMSE values for the item discrimination parameter ranged from 0.216 to 0.710 for the small separation conditions, 0.215 to 0.734 for the medium separation conditions, and 0.216 to 0.713 for the large class separation conditions for the two-class Mix3PL model solutions. Likewise, mean RMSE values for the item discrimination parameter ranged from 0.408 to 1.226 for the three-class solutions with normal class separation. The mean RMSE values for the item discrimination parameter ranged from 0.407 to 1.191 for the small separation condition, 0.367 to 1.149 for the medium separation condition, and 0.357 to 1.144 for the large class separation condition for the two-class Mix3PL model solutions. The results for item discrimination parameters of Mix3PL model are summarized in the plots in Figure 4. As expected, recovery of item discrimination parameter estimates improved as the sample size increased in the two- and three-class conditions. As shown in Figure 4, recovery was worse in for the three-class solutions compared with the results for two-class solutions under the Mix3PL model. The 10-item conditions with nonequal group sizes (NE10) produced the highest mean RMSE values and 50-item conditions with equal- and nonequal group sizes (E50 and NE50) had the lowest mean RMSE values for both two- and three-class conditions. Increase in the test length had a positive effect on the recovery of item difficulty parameter estimates. Overall, nonequal group size conditions had higher RMSE values than conditions with equal group size conditions for each test length conditions. None of the six conditions yielded mean RMSE values less than 0.10. Mean RMSE values were large for the conditions with fewer than 1,000 examinees particularly with short tests. The degree of class separation appeared to have a positive effect on the recovery of item difficulty parameters.

Comparable results in Figures 3 and 4 were observed for item difficulty and item guessing parameters of Mix3PL model (see Figures 5 and 6). However, the difference between 10-item conditions and other test lengths decreased for item difficulty for the Mix3PL model, compared with the results of Mix1PL and Mix2PL models. As for item discrimination results, the number of latent classes, sample size, test length,

the degree of class separation, and class size appeared to affect recovery of item difficulty and guessing parameters for the Mix3PL model. The conditions with larger sample sizes, longer tests, smaller number of latent classes, and equal group sizes had lower RMSEs for item difficulty and guessing parameters, compared with conditions with smaller sample sizes, shorter tests, larger number of classes and nonequal group sizes. The degree of class separation, however, appeared to differentially affect the recovery of item difficulty and guessing parameters for the two- and three-class solutions. The increase in the degree of class separation had a positive effect on the recovery of item guessing parameters in two- and three-class conditions. This was also the case with item difficulty parameters in the two-class conditions. However, the pattern was the reverse for the shorter test lengths for the three-class conditions. Minimum and maximum RMSE values of item difficulty and guessing parameter estimates for the two- and three-class solutions are also presented in Tables 7A and 8A (see Appendix).

Mean RMSEs for item difficulty parameters were above 0.10 for all conditions with all sample sizes in both two- and three-class model conditions (see Figure 5). Mean RMSE values of item guessing parameters were also above 0.10 for most of the conditions (see Figure 6). Mean RMSE values of some conditions for 30 and 50 items, however, were less than 0.10 for samples of 4000 or greater. In general, recovery of the item guessing parameter was better than that of item difficulty and item difficulty was recovered better than item discrimination for Mix3PL.

Classification Accuracy Results

Figures 7 to 9 present classification accuracy results for the Mix1PL, Mix2PL, and Mix3PL models, when model-data fit holds. Recovery of classification accuracy percentages varied between 81.22 and 99.76 for the Mix1PL two-class model for the normal class separation conditions. Classification accuracy percentages varied from 82.28 to 99.77 for the small separation conditions, from 83.32 to 99.78 for the medium separation conditions, and from 83.82 to 99.77 for the large class separation conditions for the two-class Mix1PL model. Likewise, classification accuracy percentages varied between 54.58 and 97.66 for the Mix1PL three-class model for the normal class separation condition. Classification accuracy percentages varied from 54.01 to 97.70 for the small separation condition, from 50.99 to 97.86 for the medium separation condition, and from 49.20 to 97.95 for the large class separation condition for the three-class Mix1PL model. Mean classification accuracy values were above 99% for two-class solutions, but mean classification accuracy values were lower, being above 83% for the three-class solutions. As shown in Figure 7, classification accuracy percentages for the conditions for the two-class solutions were better than for the three-class solutions. Classification accuracy percentages appeared to increase as the test length and number of examinees increased. However, the rate of increase was small for 30- and 50-item conditions. The increase was consistent for the 10-item conditions with the sample sizes larger than 400. Test conditions with 30 and 50 items yielded higher percentages than those of the 10-item conditions. In contrast to

item recovery results, conditions with nonequal group sizes had higher classification accuracy percentages than conditions with equal group sizes. The 2-class × 50-item × nonequal proportion conditions produced the highest classification accuracy percentages for the Mix1PL model for all sample sizes. Classification accuracy percentages appeared to increase as the degree of class separation increased from normal separation to large separation. Recovery of classification accuracy percentages was almost perfect (i.e., 99 percent or higher) for all two-class conditions with 50 items. Classification accuracy percentages were also very high (i.e., 98 percent or higher) for the 30-item conditions under two-class solutions. However, the percent of correct identifications was less than 90 for the two- and three-class conditions with 10 items.

Classification accuracy percentages for the Mix2PL are plotted in Figure 8. The percent of correct identifications varied between 78.85 and 99.65 for Mix2PL two-class model conditions with normal class separation. Classification accuracy percentages varied from 72.68 to 99.54 for small separation conditions, from 68.89 to 99.49 for medium separation conditions, and from 65.04 to 99.47 for large class separation for two-class Mix2PL model conditions. Likewise, classification accuracy percentages varied between 48.55 and 98.83 for Mix2PL three-class model conditions with normal class separation. Classification accuracy percentages varied from 50.85 to 98.69 for small, from 48.27 to 98.64 for medium, and from 49.53 to 98.60 for large class separations with three-class Mix2PL model conditions. Mean classification accuracy values were above 93% for two-class solution conditions, although mean classification accuracy values were lower, starting at 79% for three-class solution conditions. Classification accuracy percentages for conditions of the Mix2PL two-class solutions appear to be better than that of Mix2PL three-class solutions (see Figure 8). Classification accuracy percentages also appear to increase as the test length and number of examinees increased. However, the rate of increase of classification accuracy percentages changed only slightly as sample size changed, in the 30- and 50-item conditions. Increase was consistent for the 10-item two-class conditions for sample sizes larger than 2,000. Conditions with 30 and 50 items yielded higher percentages than the 10-item conditions. In contrast to item recovery results, conditions with nonequal group sizes produced higher classification percentages than conditions with equal mixture proportions for most of the simulation conditions. The 2-class × 50-item × nonequal proportion conditions produced the highest percentages for the Mix2PL model for all sample sizes. Classification accuracy percentages were high (at 98 percent or higher) for almost all two-class 50-item conditions. Classification accuracy percentages were also very high (above 97 percent) for 30-item conditions for the two-class solutions. The percent of correct identifications was lower, however, for the two- and three-class conditions for the 10-item tests. Classification accuracy percentages appear to decrease as the degree of class separation increased from normal to large separation.

Classification accuracy percentages for Mix3PL are plotted in Figure 9. The percent of correct identifications varied between 62.40 and 98.22 for Mix3PL two-class

model conditions with normal class separation. Classification accuracy percentages varied from 64.69 to 98.38 for the small separation conditions, from 64.08 to 98.13 for the medium separation conditions, and from 67.42 to 98.15 for the large class separation conditions for the two-class Mix3PL model. Likewise, recovery of classification accuracy percentages varied between 52.69 and 90.17 for Mix3PL three-class model conditions with normal class separation. Classification accuracy percentages varied from 51.77 to 92.31 for small class separation conditions, from 54.90 to 93.06 for medium class separation conditions, and from 54.31 to 93.34 for large class separation conditions for the three-class Mix3PL model. Mean classification accuracy values were above 86% for two-class solution conditions, but were only above 70% for three-class solution conditions. As was the case with the Mix1PL and Mix2PL, classification accuracy percentages for the conditions with two-class solutions were better than that of three-class solutions for Mix3PL model. Classification accuracy percentages appeared to increase as the test length and number of examinees increase. However, the increase rate of percentages changes slightly with sample size increase especially for 30- and 50-item conditions with sample sizes larger than 200. Increase appears to be stable for 10-item conditions with the sample sizes larger than 1,000. Test conditions with 30 and 50 items yielded higher percentages than that of 10-item conditions. Mixture proportions, however, appeared to differ depending on the recovery of item difficulty parameters. The 2-class $\times$ 50-item $\times$ equal proportion conditions produced the highest percentages for the Mix3PL model for all sample sizes. Classification accuracy percentages were relatively good for most of the two-class conditions (mostly at or above 94 percent) and three-class conditions (mostly at or above 80 percent) for 50-item tests. Classification accuracy percentages were also very high for 30-item tests for the two-class conditions (mostly at or above 90 percent) and three-class conditions (mostly at or above 80 percent). However, the percent of correct identifications for 10-item test conditions were lower, at around 60 percent for the two-class solutions, and around 70 percent for the three-class conditions. As was the case with Mix2PL, classification accuracy percentages appear to decrease as the degree of class separation increase from normal to large.

When all models were compared, the highest classification accuracy percentages were observed for the Mix1PL model for the two-class solutions with a mean across all conditions of 95.08 percent. The models with the next highest classification percentages were for the two-class Mix2PL ($M = 94.24$), two-class Mix3PL ($M = 87.08$), three-class Mix1PL ($M = 84.32$), three-class Mix2PL ($M = 81.44$), and three-class Mix3PL ($M = 71.46$).

A Linear Model Analysis of Simulation Results

Mean RMSE and classification accuracy results were also summarized using a linear model. Effects of each of the conditions were evaluated using a factorial ANOVA for the RMSE and classification accuracy percentages. The partial eta-squared and F-values from the factorial ANOVA are presented in Table 3 for each main effect and 2-

Table 3. Partial Eta-Squared Values for Main Effects and 2-Way Interactions of Simulation Conditions.

	α		β		γ		CA%	
Factor	*F*	partial-η^2	*F*	partial-η^2	*F*	partial-η^2	*F*	partial-η^2
N	561.291	.857**	178.03	.552**	5267.97	.992**	120.820	.456**
k	965.255	.674**	24.691	.033**	9665.68	.979**	3,418.561	.826**
P	10.633	.011**	23.414	.016**	42.117	.091**	32.572	.022*
M	2747.64	.747**	625.976	.464**	—	—	850.904	.541**
C	1783.92	.657**	57.994	.039**	.000	.000	3,514.560	.709**
S	5.719	.018**	368.892	.434**	695.927	.832**	9.691	.020**
N × *k*	1.116	.023	1.467	.020	63.547	.750**	3.779	.050**
N × *P*	0.470	.005	0.029	.000	2.914	.064*	0.265	.002
N × *M*	21.249	.186**	5.100	.066**	—	—	8.471	.105**
N × *C*	23.913	.204**	0.083	.001	.000	.000	21.936	.132**
N × *S*	0.379	.012	5.863	.109**	9.277	.397**	0.162	.003
k × *P*	2.957	.006	2.619	.004	131.098	.383**	9.005	.012**
k × *M*	217.997	.318**	94.734	.208**	—	—	23.836	.062**
k × *C*	35.944	.072**	11.873	.016**	.000	.000	220.372	.234**
k × *S*	1.071	.007	10.462	.042**	13.844	.164**	0.374	.002
P × *M*	7.077	.008*	4.319	.006*	—	—	26.840	.036**
P × *C*	6.264	.007*	42.818	.029**	.000	.000	16.511	.011**
P × *S*	0.281	.001	1.179	.002	21.403	.132**	1.557	.003
M × *C*	97.144	.094**	89.229	.058**	—	—	41.036	.054**
M × *S*	4.661	.015**	50.387	.173**	—	—	2.454	.010*
C × *S*	0.418	.001	1.691	.004	.000	.000	1.853	.004

Note. *N* = sample size, *k* = test length (i.e., number of items), *p* = mixture proportions, *M* = model type, *C* = number of classes. α = item discrimination; β = item difficulty; γ = item guessing, CA = classification accuracy. *$p < .05$. **$p < .01$.

way interactions. As can be seen, number of examinees (*N*), test length (*k*), mixture proportions (*P*), class separation (*S*), the number of latent classes (*C*), and model type (*M*) significantly affected a part of the variation in both the RMSE and classification accuracy percentages. However, the number of latent classes (*C*) did not explained variation only in the item guessing parameter of Mix3PL conditions.

Based on partial eta-squared values, sample size (*N*) and test length (*k*) were the most influential factors on RMSE and classification accuracy percentages for each item parameter. Model type had also a large effect on the results. The least influential factor was the mixture proportions.

In addition to the interaction between number of class and two factors: mixture proportion (*P* × *C*) and test length (*k* × *C*), interaction between model type and other factors including sample size (*N* × *M*), test length (*k* × *M*), number of class (*M* × *C*), mixture proportions (*P* × *M*) and latent class separation (*M* × *S*) affected both the RMSE values and classification accuracy parameters. Mean RMSE values for item guessing parameter were also significantly affected by other two-way

interactions including sample size (N) × test length (k), sample size (N) × mixture proportion (P), sample size (N) × class separation (S), test length (k) × mixture proportion (P), and test length (k) × class separation (S). These results suggest that interactions of test length with other factors such as sample size, model type and number of latent classes may affect model parameter estimates.

Discussion

Although mixture IRT models have been shown to be useful in educational and psychological measurement, relatively little research has been reported on the effects of practical testing conditions such as sample size, test length, number of latent classes, the degree of latent class separation, and mixture proportions on IRT model parameter estimates or classification accuracy. In this study, a Monte Carlo simulation study was conducted to examine the effects of these testing conditions on item parameter recovery and classification accuracy for three dichotomous mixture IRT models. RMSE values were computed between generating parameters and the parameter estimates. Effects on class assignment were also assessed using the percentage of classification accuracy.

The effect of sample size on mixture IRT model parameter estimates has only been partially investigated in previous studies. The sample size conditions in this study included a range of conditions from 100 to 5,000 simulated examinees, whereas previous research has only reported two or three different sample size conditions. Our findings demonstrated that recovery of item parameters and classification accuracy was better with an increase in sample size and test length. These results are consistent with simulation results in S. J. Cho et al. (2013) and Preinerstorfer and Formann (2012). Item parameter recovery and class assignment were better for the two-class models than the three-class models. In addition, item parameter recovery was better for equal class sizes, but the reverse was the case for classification accuracy, with models with larger class separations producing less accurate results.

Recovery of item parameter estimates and classification accuracy percentages was better with larger samples of examinees. This was consistent with results from previous studies (S. J. Cho et al., 2013; Preinerstorfer & Formann, 2012). In this study, however, we investigated smaller sample sizes to determine where smaller sample sizes might be expected to affect recovery. Thus, we report on recovery results for each model for sample sizes from very small (i.e., 100) to large (i.e., 5,000). Results of this study suggest that the more complex mixture IRT models (e.g., the Mix3PL model) generally require larger sample sizes to yield accurate estimates. Furthermore, results of this study also suggest that the simpler mixture IRT models (e.g., the MRM and the Mix1PL) can typically be estimated accurately with smaller sample sizes. The findings of this study also suggest that sample size should be at least 2,000 to obtain stable item difficulty estimates for the Mix1PL. Furthermore, the need for larger sample size increased as model complexity increased. For the Mix2PL model, at least 3000 examinees were needed for accurate recovery of item difficulty

estimates and the larger sample sizes were needed for recovery of item discrimination. For the Mix3PL model, a sample size of 5,000 was needed to obtain RMSE values lower than 0.10 for all item parameters. That is, results suggest that accuracy of recovery of item parameters estimates improved as sample sizes increased. The RMSE tended to be slightly larger for the Mix2PL and Mix3PL models than the Mix1PL model. That is, as the number of item parameters increased, recovery tended to decrease for a given sample size.

Findings also suggested some increase in classification accuracy percentages with an increase in sample size, although this increase was relatively small for 30- and 50-item conditions. Classification accuracy percentages, however, did not appear to be affected by sample sizes larger than 1,000. Effects of sample size on classification accuracy with 10 items were greater than for the 30- and 50-item conditions. The classification accuracy percentages were also positively affected, however, by an increase in test length and a decrease in number of latent classes. Finally, classification accuracy rates were higher, when the class sizes were not equal.

In summary, consistent with previous research (S. J. Cho et al., 2013; W. H. Finch & French, 2012; Li et al., 2009), item parameter recovery and classification accuracy were better with an increase in test length, although classification accuracy percentages varied, as a function of test length. Recovery results for mixture IRT models in this study were better with more than 10 items. This was also the case for recovery of item parameters and class assignment.

The results of this simulation study also showed that two-class mixture IRT models were recovered better than three-class as were recovery of item difficulty and discrimination parameters. That is, more classes appeared to result in less accurate recovery of item parameters and less accurate classification for all models. The increase in number of latent classes for a given sample size condition was also associated with a decrease in sample size in each class. That is, for a given sample size, more latent classes result in smaller samples in latent classes. Thus, decreased accuracy of recovery of item parameters and the lower classification accuracy are likely due, at least in part, to the smaller number of individuals in the latent classes, as suggested by W. H. Finch and French (2012). Classification accuracy rates, in other words, were higher for two-class solution conditions. Results of this study were consistent with results reported in Li et al. (2009).

Another important finding obtained is the effect of item parameter recovery was better for equal class sizes but less so compared with the nonequal class sizes. Results also suggested that studies with smaller sample sizes or shorter test lengths may do well to consider simpler models such as the Mix1PL and MRM.

Results of this study showed that the increase in latent class separation had a negative effect on both the recovery of item parameters and classification accuracy. When latent class separation was large, all three models produced less accurate recovery of generating parameters. Overall, better recovery was obtained with the normal separation condition in which each class was generated to have the same ability distribution

$N(0, 1)$. Only item discrimination parameter estimates of some conditions were found to show better results with larger class separation conditions.

It should be noted that estimation of models did not run smoothly for all conditions. With respect to using mixture IRT models with relatively small samples (i.e., of 500 or less), MLR estimation had difficulty reaching convergence for many of the conditions in the study. While stable results were obtained for conditions with more than 500 cases, more than 100 data sets were generated in conditions with small samples and only converged data sets included in this study. This problem was particularly evident for a conditions with three latent classes and for the more complex Mix3PL models for smaller sample sizes. For the small sample size conditions with complex models, the variance/covariance matrices of the mixture IRT models included some unidentified values (e.g., caused by nonpositive definite matrices) and multiple maxima problems. The data sets generated for shorter test conditions also had difficulty reaching convergence compared with data sets with the longer tests.

Future research on mixture IRT models might consider examining the effects of the appropriateness of nonnormal distributions on accuracy of recovery of item parameter estimates. In this regard, it is possible that the effects of sample size may vary, when the distribution of ability is nonnormal. In addition, examining the recovery of item parameters for different mixture IRT models, for example, multilevel mixture IRT models, would also be useful.

Appendix

Table 1A. Generating Item Parameter Values for 30-Item Conditions.

Item	β_1	β_2	β_3	α_1	α_2	α_3	γ_1	γ_2	γ_3
1	–0.680	1.079	0.759	1	2	1	0.10	0.25	0.20
2	–0.310	1.463	1.105	1	2	1	0.10	0.25	0.20
3	–1.115	1.004	1.328	1	2	1	0.10	0.25	0.20
4	–1.853	1.242	0.917	1	2	1	0.10	0.25	0.20
5	–1.714	0.665	0.467	1	2	1	0.10	0.25	0.20
6	–0.256	0.175	–1.158	1	2	1	0.10	0.25	0.20
7	–0.457	1.952	0.522	1	2	1	0.10	0.25	0.20
8	–0.017	1.516	–0.729	1	2	1	0.10	0.25	0.20
9	–0.057	0.326	–0.264	1	2	1	0.10	0.25	0.20
10	–0.918	0.574	–0.986	1	2	1	0.10	0.25	0.20
11	–0.577	0.952	–1.819	1	2	2	0.20	0.20	0.25
12	–1.120	0.483	–0.763	1	2	2	0.20	0.20	0.25
13	–1.773	1.009	–1.119	1	2	2	0.20	0.20	0.25
14	–0.381	0.138	–1.632	1	2	2	0.20	0.20	0.25
15	–0.487	–1.203	–0.637	1	2	2	0.20	0.20	0.25
16	–0.804	–0.303	–1.294	2	1	2	0.20	0.20	0.25
17	–1.305	–1.025	–0.599	2	1	2	0.20	0.20	0.25
18	–0.301	–1.894	–0.971	2	1	2	0.20	0.20	0.25
19	1.155	–1.460	–0.717	2	1	2	0.20	0.20	0.25
20	0.722	–0.826	–0.559	2	1	2	0.20	0.20	0.25
21	1.725	–1.710	0.471	2	1	1	0.25	0.10	0.10
22	1.508	–0.942	1.268	2	1	1	0.25	0.10	0.10
23	1.144	–0.481	0.679	2	1	1	0.25	0.10	0.10
24	0.916	–1.133	0.113	2	1	1	0.25	0.10	0.10
25	0.893	–0.250	1.898	2	1	1	0.25	0.10	0.10
26	1.035	–1.010	0.197	2	1	1	0.25	0.10	0.10
27	0.924	–1.685	1.123	2	1	1	0.25	0.10	0.10
28	0.795	–0.577	0.819	2	1	1	0.25	0.10	0.10
29	0.927	–0.270	1.038	2	1	1	0.25	0.10	0.10
30	0.171	–0.493	1.558	2	1	1	0.25	0.10	0.10

Table 2A. Generating Item Parameter Values for 50-Item Conditions.

Item	β_1	β_2	β_3	α_1	α_2	α_3	γ_1	γ_2	γ_3	Item	β_1	β_2	β_3	α_1	α_2	α_3	γ_1	γ_2	γ_3
1	−1.229	0.088	0.861	1	2	1	0.10	0.25	0.20	26	0.003	−1.390	−0.484	2	1	2	0.20	0.20	0.25
2	−1.400	1.065	0.328	1	2	1	0.10	0.25	0.20	27	0.012	−0.247	−0.540	2	1	2	0.20	0.20	0.25
3	−1.206	1.154	0.307	1	2	1	0.10	0.25	0.20	28	0.767	−1.334	−0.307	2	1	2	0.20	0.20	0.25
4	−0.647	0.652	0.386	1	2	1	0.10	0.25	0.20	29	0.828	−0.175	−0.276	2	1	2	0.20	0.20	0.25
5	−1.126	0.993	0.612	1	2	1	0.10	0.25	0.20	30	0.810	−1.204	−0.416	2	1	2	0.20	0.20	0.25
6	−0.364	1.781	0.112	1	2	1	0.10	0.25	0.20	31	1.524	−0.716	−0.524	2	1	1	0.20	0.20	0.25
7	−0.180	1.819	1.757	1	2	1	0.10	0.25	0.20	32	1.995	−1.269	−0.419	2	1	1	0.20	0.20	0.25
8	−0.395	1.906	0.591	1	2	1	0.10	0.25	0.20	33	0.573	−1.672	−1.699	2	1	1	0.20	0.20	0.25
9	−0.952	0.367	1.273	1	2	1	0.10	0.25	0.20	34	1.821	−1.282	−0.296	2	1	1	0.20	0.20	0.25
10	−1.364	0.195	1.565	1	2	1	0.10	0.25	0.20	35	0.971	−0.035	−0.492	2	1	1	0.20	0.20	0.25
11	−0.498	0.042	1.104	1	2	1	0.10	0.25	0.20	36	0.579	−1.305	−0.642	2	1	1	0.25	0.10	0.10
12	−1.467	0.942	1.119	1	2	1	0.10	0.25	0.20	37	0.520	−0.370	−1.337	2	1	1	0.25	0.10	0.10
13	−0.402	0.952	0.238	1	2	1	0.10	0.25	0.20	38	0.905	−0.896	−0.779	2	1	1	0.25	0.10	0.10
14	−0.507	1.179	−0.629	1	2	1	0.10	0.25	0.20	39	1.110	−0.577	1.369	2	1	1	0.25	0.10	0.10
15	−0.112	0.559	−1.481	1	2	1	0.10	0.25	0.20	40	0.443	−0.545	0.096	2	1	1	0.25	0.10	0.10
16	−0.848	0.355	−1.082	1	2	1	0.20	0.20	0.25	41	0.553	−0.358	0.218	2	1	1	0.25	0.10	0.10
17	−1.240	0.569	−1.084	1	2	1	0.20	0.20	0.25	42	1.475	−0.458	1.414	2	1	1	0.25	0.10	0.10
18	−0.852	0.317	−0.600	1	2	1	0.20	0.20	0.25	43	0.033	−0.844	1.630	2	1	1	0.25	0.10	0.10
19	−0.590	1.218	−1.863	1	2	1	0.20	0.20	0.25	44	0.638	−0.025	1.720	2	1	1	0.25	0.10	0.10
20	−1.112	1.14	−0.345	1	2	1	0.20	0.20	0.25	45	1.838	−0.822	1.000	2	1	1	0.25	0.10	0.10
21	−0.845	0.903	−0.613	1	2	2	0.20	0.20	0.25	46	1.679	−0.213	0.018	2	1	1	0.25	0.10	0.10
22	−1.646	1.512	−1.972	1	2	2	0.20	0.20	0.25	47	1.404	−0.123	1.699	2	1	1	0.25	0.10	0.10
23	−0.672	0.386	−1.137	1	2	2	0.20	0.20	0.25	48	1.265	−1.441	0.843	2	1	1	0.25	0.10	0.10
24	−1.534	1.271	−0.342	1	2	2	0.20	0.20	0.25	49	1.703	−0.853	1.052	2	1	1	0.25	0.10	0.10
25	−0.663	1.598	−0.708	1	2	2	0.20	0.20	0.25	50	1.145	−1.568	1.193	2	1	1	0.25	0.10	0.10

Table 3A. Mean RMSE Ranges for Item Difficulty Parameter Estimates of Mix1PL Model.

	β							
	Normal		Small		Medium		Large	
Condition	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Two-class-E10	0.065	0.461	0.287	0.547	0.423	0.640	0.484	0.702
Two-class-NE10	0.076	0.511	0.304	0.629	0.440	0.720	0.499	0.806
Two-class-E30	0.053	0.359	0.265	0.454	0.399	0.548	0.478	0.606
Two-class-NE30	0.058	0.404	0.276	0.520	0.410	0.627	0.490	0.690
Two-class-E50	0.051	0.352	0.275	0.455	0.407	0.559	0.484	0.629
Two-class-NE50	0.058	0.389	0.284	0.511	0.414	0.624	0.491	0.700
Three-class-E10	0.122	0.781	0.373	0.823	0.550	0.905	0.582	0.774
Three-class-NE10	0.133	0.806	0.367	0.819	0.537	0.884	0.583	0.834
Three-class-E30	0.070	0.579	0.352	0.639	0.537	0.721	0.550	0.731
Three-class-NE30	0.074	0.621	0.348	0.673	0.533	0.759	0.546	0.751
Three-class-E50	0.064	0.449	0.351	0.548	0.533	0.680	0.553	0.696
Three-class-NE50	0.066	0.495	0.345	0.568	0.528	0.664	0.550	0.676

Note. RMSE = root mean square error.

Table 4A. Mean RMSE Ranges for Item Discrimination Parameter Estimates of Mix2PL Model.

	α							
	Normal		Small		Medium		Large	
Condition	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Two-class-E10	0.154	0.961	0.141	0.915	0.141	0.959	0.147	0.918
Two-class-NE10	0.174	0.901	0.177	0.870	0.185	0.817	0.177	0.793
Two-class-E30	0.087	0.569	0.087	0.587	0.087	0.612	0.088	0.648
Two-class-NE30	0.097	0.652	0.097	0.651	0.098	0.670	0.098	0.669
Two-class-E50	0.081	0.525	0.080	0.517	0.081	0.530	0.081	0.530
Two-class-NE50	0.090	0.587	0.089	0.577	0.091	0.578	0.091	0.555
Three-class-E10	0.337	1.140	0.252	1.121	0.288	1.161	0.324	1.123
Three-class-NE10	0.303	1.169	0.263	1.121	0.369	1.175	0.398	1.171
Three-class-E30	0.109	0.783	0.106	0.780	0.109	0.780	0.112	0.787
Three-class-NE30	0.115	0.818	0.112	0.816	0.114	0.818	0.116	0.824
Three-class-E50	0.149	0.874	0.165	0.903	0.184	0.956	0.097	0.649
Three-class-NE50	0.099	0.647	0.101	0.628	0.102	0.668	0.103	0.690

Note. RMSE = root mean square error.

Table 5A. Mean RMSE Ranges for Item Difficulty Parameter Estimates of Mix2PL Model.

	β							
	Normal		Small		Medium		Large	
Condition	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Two-class-E10	0.103	0.671	0.308	0.709	0.443	0.757	0.529	0.799
Two-class-NE10	0.124	0.698	0.345	0.762	0.489	0.822	0.558	0.878
Two-class-E30	0.059	0.377	0.269	0.453	0.403	0.545	0.485	0.619
Two-class-NE30	0.065	0.457	0.281	0.575	0.414	0.667	0.493	0.713
Two-class-E50	0.057	0.365	0.277	0.455	0.411	0.543	0.490	0.604
Two-class-NE50	0.064	0.409	0.290	0.532	0.424	0.635	0.502	0.692
Three-class-E10	0.231	0.913	0.396	0.917	0.556	0.918	0.663	0.929
Three-class-NE10	0.205	0.947	0.372	0.928	0.541	0.969	0.633	0.934
Three-class-E30	0.078	0.607	0.348	0.626	0.527	0.686	0.636	0.708
Three-class-NE30	0.084	0.632	0.343	0.650	0.514	0.686	0.607	0.722
Three-class-E50	0.120	0.815	0.397	0.820	0.589	0.861	0.644	0.710
Three-class-NE50	0.076	0.498	0.346	0.524	0.516	0.637	0.626	0.756

Note. RMSE = root mean square error.

Table 6A. Mean RMSE Ranges for Item Discrimination Parameter Estimates of Mix3PL Model.

	α							
	Normal		Small		Medium		Large	
Condition	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Two-class-E10	0.538	0.794	0.533	0.701	0.541	0.650	0.569	0.659
Two-class-NE10	0.536	0.801	0.629	0.710	0.619	0.734	0.479	0.713
Two-class-E30	0.264	0.629	0.269	0.603	0.253	0.607	0.250	0.590
Two-class-NE30	0.274	0.626	0.288	0.586	0.318	0.554	0.330	0.542
Two-class-E50	0.244	0.648	0.216	0.591	0.215	0.564	0.216	0.557
Two-class-NE50	0.264	0.644	0.251	0.600	0.263	0.553	0.282	0.553
Three-class-E10	0.543	1.143	0.568	1.120	0.571	1.085	0.596	1.036
Three-class-NE10	0.541	1.145	0.607	1.111	0.584	1.082	0.607	1.072
Three-class-E30	0.408	1.053	0.421	0.995	0.418	0.974	0.421	0.954
Three-class-NE30	0.421	1.045	0.422	1.028	0.430	0.989	0.461	1.001
Three-class-E50	0.519	1.213	0.407	1.140	0.367	1.149	0.357	1.126
Three-class-NE50	0.602	1.226	0.531	1.191	0.536	1.137	0.485	1.144

Note. RMSE = root mean square error.

Table 7A. Mean RMSE Ranges for Item Difficulty Parameter Estimates of Mix3PL Model.

	β							
	Normal		Small		Medium		Large	
Condition	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Two-class-E10	0.485	1.046	0.563	1.027	0.640	1.023	0.717	1.046
Two-class-NE10	0.485	1.000	0.727	0.997	0.788	1.015	0.768	0.980
Two-class-E30	0.261	0.899	0.406	0.931	0.518	0.959	0.584	0.961
Two-class-NE30	0.255	0.966	0.445	0.998	0.629	1.030	0.718	1.073
Two-class-E50	0.220	0.792	0.363	0.846	0.480	0.883	0.557	0.902
Two-class-NE50	0.226	0.823	0.393	0.881	0.537	0.898	0.626	0.944
Three-class-E10	0.497	0.888	0.480	0.879	0.464	0.849	0.478	0.829
Three-class-NE10	0.509	0.916	0.475	0.886	0.478	0.875	0.475	0.836
Three-class-E30	0.385	1.040	0.476	1.024	0.586	1.013	0.625	1.008
Three-class-NE30	0.412	1.050	0.490	1.050	0.583	1.029	0.612	1.011
Three-class-E50	0.417	1.087	0.499	1.098	0.607	1.095	0.684	1.091
Three-class-NE50	0.520	1.087	0.548	1.084	0.624	1.077	0.632	1.098

Note. RMSE = root mean square error.

Table 8A. Mean RMSE Ranges for Item Guessing Parameter Estimates of Mix3PL Model.

	γ							
	Normal		Small		Medium		Large	
Condition	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
Two-class-E10	0.164	0.252	0.172	0.302	0.173	0.307	0.172	0.318
Two-class-NE10	0.153	0.250	0.158	0.282	0.170	0.287	0.181	0.294
Two-class-E30	0.099	0.255	0.111	0.272	0.108	0.285	0.102	0.285
Two-class-NE30	0.097	0.261	0.110	0.284	0.122	0.297	0.126	0.305
Two-class-E50	0.079	0.229	0.085	0.246	0.089	0.257	0.093	0.263
Two-class-NE50	0.082	0.228	0.089	0.246	0.094	0.258	0.099	0.268
Three-class-E10	0.164	0.252	0.172	0.302	0.173	0.307	0.172	0.318
Three-class-NE10	0.153	0.250	0.158	0.282	0.170	0.287	0.181	0.294
Three-class-E30	0.099	0.255	0.111	0.272	0.108	0.285	0.102	0.285
Three-class-NE30	0.097	0.261	0.110	0.284	0.122	0.297	0.126	0.305
Three-class-E50	0.079	0.229	0.085	0.246	0.089	0.257	0.093	0.263
Three-class-NE50	0.082	0.228	0.089	0.246	0.094	0.258	0.099	0.268

Note. RMSE = root mean square error.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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

ORCID iD

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

Supplemental Material

Supplemental material for this article is available online.

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*(3), 313–332.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, *40*(6), 1235–1245.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling approach. *Applied Psychological Measurement*, *22*, 153–169.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, *26*(4), 381–409.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348.

Cassiday, K. R., Cho, Y., & Harring, J. R. (2021). A comparison of label switching algorithms in the context of growth mixture models. *Educational and Psychological Measurement*, *81*(4), 668–697.

Cho, H. J., Lee, J., & Kingston, N. (2012). Examining the effectiveness of test accommodation using DIF and a mixture IRT model. *Applied Measurement in Education*, *25*(4), 281–304.

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.

Cho, S. J., Cohen, A. S., Kim, S. H., & Bottge, B. (2010). Latent transition analysis with a mixture item response theory measurement model. *Applied Psychological Measurement*, *34*(7), 483–504.

Cho, Y. (2014). *The mixture distribution polytomous Rasch model used to account for response styles on rating scales: A simulation study of parameter recovery and classification accuracy* (dissertation abstracts international 75). http://hdl.handle.net/1903/14511

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148.

Cohen, A. S., Bottge, B. A., & Wells, C. S. (2001). Using item response theory to assess effects of mathematics instruction in special populations. *Exceptional Children*, *68*(1), 23–44.

Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research & Practice*, *20*(4), 225–233.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement*, *34*(4), 267–285.

DeMars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: How accurately can we detect who is responding differentially? *Educational and Psychological Measurement*, *71*(4), 597–616.

Finch, H., & French, B. F. (2019). A comparison of estimation techniques for IRT models with small samples. *Applied Measurement in Education*, *32*(2), 77–96.

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*(1), 167–178.

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*(4), 432–457.

Glück, J., Machat, R., Jirasko, M., & Rollett, B. (2002). Training-related changes in solution strategy in a spatial test: An application of item response models. *Learning and Individual Differences*, *13*(1), 1–22.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large scale latent variable analyses in M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.

Hamner, B., Frasco, M., & LeDell, E. (2018). *Package ''Metrics''* (Version 0.1.4). https://cran.r-project.org/web/packages/Metrics/Metrics.pdf

Huang, H. Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology*, *7*, Article 1706.

Jiao, H., Macready, G., Liu, J., & Cho, Y. (2012). A mixture Rasch model–based computerized adaptive test for latent class identification. *Applied Psychological Measurement*, *36*(6), 469–493.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, *27*(4), 307–327.

Kutscher, T., Eid, M., & Crayen, C. (2019). Sample size requirements for applying mixed Polytomous item response models: Results of a Monte Carlo simulation study. *Frontiers in Psychology*, *10*, 2494.

Lee, S., Han, S., & Choi, S. W. (2021). DIF detection with zero-inflation under the factor mixture modeling framework. *Educational and Psychological Measurement*. https://www.sciencegate.app/document/10.1177/00131644211028995

Lee, W. Y., Cho, S. J., & Sterba, S. K. (2018). Ignoring a multilevel structure in mixture item response models: Impact on parameter recovery and model selection. *Applied Psychological Measurement*, *42*(2), 136–154.

Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*(5), 353–373.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, *32*(8), 611–631.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, *45*(6), 975–999.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195–215.

Nye, C. D., Joo, S. H., Zhang, B., & Stark, S. (2020). Advancing and evaluating IRT model data fit Indices in organizational research. *Organizational Research Methods*, *23*(3), 457–486.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333.

Ölmez, İ. B., & Cohen, A. S. (2018). A mixture partial credit analysis of math anxiety. *International Journal of Assessment Tools in Education*, *5*(4), 611–630.

Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 251–262.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282.

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

Sen, S. (2016). Applying the mixed Rasch model to the Runco ideational behavior scale. *Creativity Research Journal*, *28*(4), 426–434.

Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177–191.

Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models* [Technical report]. MRC Biostatistics Uni.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 13–30). Academic Press.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, *27*(1), 27-51.

Wetzel, E., Böhnke, J. R., & Rose, N. (2016). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement*, *76*(2), 304–324.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116.