International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 1013–1029
https://doi.org/10.21449/ijate.1164590
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Investigation of the effect of parameter estimation and classification accuracy in mixture IRT models under different conditions

Fatima Munevver Saatcioglu[1,*], **Hakan Yavuz Atar**[2]

[1]Ankara Yildirim Beyazit University, Rectorate, Ankara, Turkiye
[2]Gazi University, Faculty of Education, Department of Educational Sciences, Ankara, Turkiye

ARTICLE HISTORY
Received: Aug. 19, 2022
Revised: Dec. 07, 2022
Accepted: Dec. 18, 2022

Keywords:
Mixture Item Response Theory Models,
Maximum Likelihood Estimation,
Item Parameter Recovery,
Classification Accuracy,
Missing Data,
Latent Class.

Abstract: This study aims to examine the effects of mixture item response theory (IRT) models on item parameter estimation and classification accuracy under different conditions. The manipulated variables of the simulation study are set as mixture IRT models (Rasch, 2PL, 3PL); sample size (600, 1000); the number of items (10, 30); the number of latent classes (2, 3); missing data type (complete, missing at random (MAR) and missing not at random (MNAR)), and the percentage of missing data (10%, 20%). Data were generated for each of the three mixture IRT models using the code written in R program. *MplusAutomation* package, which provides the automation of R and Mplus program, was used to analyze the data. The mean RMSE values for item difficulty, item discrimination, and guessing parameter estimation were determined. The mean RMSE values as to the Mixture Rasch model were found to be lower than those of the Mixture 2PL and Mixture 3PL models. Percentages of classification accuracy were also computed. It was noted that the Mixture Rasch model with 30 items, 2 classes, 1000 sample size, and complete data conditions had the highest classification accuracy percentage. Additionally, a factorial ANOVA was used to evaluate each factor's main effects and interaction effects.

1. INTRODUCTION

Tests are widely used in different contexts such as education, psychology, industry, and health. In educational and psychological fields, test results are preferred for various purposes such as selecting individuals, following their development, or evaluating the efficiency of education systems. A growing awareness of the importance and the impact of testing has led to designing better tests and developing statistical methods used for the analysis of test scores. Item Response Theory (IRT) models are among the most commonly used models in various testing settings. Although IRT models have many advantages, they have strict assumptions such as unidimensionality, homogeneity population, local independence, and the invariance of item parameters (Embretson & Reise, 2000; Hambleton et al., 1991). The advantages of IRT models depend on the validity of the model whose assumptions are to be met. Traditional IRT models assume that data are drawn from a single homogeneous population. However, it may not always be possible because population may include two or more subpopulations that consist of different

*CONTACT: F. Munevver Saatcioglu fmyigiter@gmail.com , Ankara Yildirim Beyazit University, Rectorate, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

latent classes. Mixture IRT models assume that the overall population includes multiple latent classes that can be identified based on the item response patterns (Rost, 1990). In this case, the mixture IRT modeling approach is used. In social science research, there have been many studies that use mixture IRT models (Alexeev et al., 2011; Cohen et al., 2005; De Ayala & Santiago, 2017; Finch & French, 2012; Maij-de Meij et al., 2008; Lee, 2012; Oliveri et al., 2014; Sen, 2016; Zhang et al., 2015). The three-parameter Mixture IRT model including item parameters and the guessing parameter for each class is shown as the following equation:

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

In equation (1), $g=(1,2,..,G)$ indicates latent class membership, (β_{ig}), (α_{ig}), and (Υ_{ig}) represent the difficulty, discrimination, and guessing parameters, respectively for item *i*, (θ_{jg}) denotes the ability parameter for individual *j* in class g , and π_g indicates the mixing proportion of individuals in a class. The probability that each individual belongs to one latent class and the mixing proportion of individuals in each class is estimated with the (π_g), $\sum_{i=1}\pi_g=1$ and 0 $\leq\pi_g\leq1$ restriction (Rost, 1990). When the guessing parameter is equal to zero, the two-parameter mixture IRT model; with the assumption that the guessing parameter is equal to zero and the item discrimination parameter is equal to 1, the Mixture Rasch model can be obtained.

Much of the current research has focused on the Rasch and 2PL version of mixture IRT models, while there is a relatively small body of literature on the Mixture 3PL model (Cho, Cohen & Kim, 2013; Choi et al., 2020; Li et al., 2009).

When the Mixture IRT literature is examined, the sample size, the number of items, and the number of latent classes appear to affect parameter estimates of the Mixture IRT models. For example, Preinerstorfer and Formann (2012) indicated that increasing the sample size (500, 1000, 2500) and the number of items (10, 15, 25, 40) leads to higher accuracy in estimating the parameters of the mixture Rasch model. Moreover, Li et al. (2009) found that recovery of item parameters in mixture models such as the one-parameter logistic (1PL), the two-parameter logistic (2PL), and the three-parameter logistic (3PL) differed based on the sample sizes (600, 1200); the number of latent classes (1, 2, 3, 4); and the number of items (6, 15, 30). When the number of latent classes increased, the mean root mean square error (RMSE) values increased for item difficulty and disrimination parameters. Also, according to the study of Li et. al (2009), the mean RMSE values decreased as the sample size and the number of items increased. The classification accuracy increased with an increasing number of items. Different sets of sample size, number of items, and number of classes that have been used in the mixture IRT models in previous studies can be seen in the review study by Sen and Cohen (2019). The present study focuses specifically on examining the effects of factors on the estimation of item parameters and classification accuracy for mixture IRT models including 1PL, 2PL and 3PL.

Also, it is suggested that the data set should be examined in terms of missing data so that the latent variables which the tests aim to measure can be obtained (Little & Rubin, 1987). Missing data in the response patterns cause negative situations such as bias, higher standard errors in parameter estimations, and lower power of a test (De Ayala et al., 2001; Finch, 2008; Hohensinn & Kubinger, 2011; Pohl et al., 2014). At this point, it would be beneficial to determine the percentage of missing data and the mechanism of the missing data type before analyzing the data. Also, there is no study with missing data and 3PL mixture IRT models in the literature. In the context of the findings to be obtained from this study, it is therefore thought that the research is important in terms of making extensive and detailed comments on the error values and

classification accuracy obtained as a result of the mixture 3PL model and the missing data type and missing data percentage factors.

Another significance of the research is examining the RMSE and bias values of the parameter estimations obtained from the mixture IRT models, which is important in terms of evaluating the performance of the mixture IRT models in different conditions and determining which model has less errors in the determined conditions. The findings to be obtained in this direction are considered important in terms of providing information and guiding the practitioners in terms of which model would be appropriate to choose according to their own conditions in their studies.

In line with these purposes, this study tries to answer the following questions:

1) How do the mean RMSE values obtained through parameter estimations change based on the sample size, the number of items, the estimation model, the number of classes, the percentage of missing data and the missing data type factors?

2) How does the interaction effect of the variables considered change according to the mean RMSE values obtained as a result of parameter estimations?

3) How does the classification accuracy obtained from the combination of the factors change?

2. METHOD

In this study, the factors for simulation conditions were designed to investigate the effects of the model, number of latent classes, number of items, sample size, model missing data type and missing data rate on the estimates of mixture IRT model parameters and classification accuracy. The simulation conditions for this study are as follows: three Mixture IRT models (Rasch, 2PL, 3PL); number of latent classes (2, 3); number of items (10, 30); sample sizes (600, 1000); missing data mechanisms ((complete data, missing at random (MAR), missing at not random (MNAR), and missing data percentages (10%, 20%). Overall, 144 conditions were simulated in this study. One hundred replications were generated for each condition. All data sets were analyzed for each of the mixture IRT models with the computer program Mplus version 8.5 (Muthe´n & Muthe´n, 1998-2020).

2.1. Simulation Conditions

2.1.1. *Number of classes*

The examinees have different response patterns on items and according to these different patterns, they are assigned to different latent classes. This situation enables estimating group-specific parameters for latent classes in mixture IRT models. According to the study conducted by Sen and Cohen (2019), the number of latent classes used in the studies ranges from one to ten. However, according to the results of the model-data fit studies, it is stated that the data generally fit the mixture IRT model with two or three latent classes (Finch & French, 2012; Park et al., 2016). Therefore, in this study, the conditions for the number of classes were determined as two and three to identify poor, average, and good performing individuals (Li et al., 2009).

2.1.2. *Number of items*

The number of items has been one of the manipulated variables in various simulation studies in the existing literature. The study conducted by Sen and Cohen (2019) shows that the number of items used in previous studies varies between 4 and 470 (Cho et al., 2012; Jilke et al., 2015). In this research, item numbers were taken as 10 and 30 as reported in Lee (2012) to generate different profile of latent classes (poor, average and good performing) according to item parameter values.

2.1.3. *Distribution of item and ability parameters*

Data were generated for each mixture IRT model (i.e., Rasch, 2PL, 3PL) using R program (R Core Team, 2020). The distributions of ability and item parameters were generated to be the same for each model. Then, class-specific item parameters were generated for each model and item parameter values for the classes were obtained (see Table 1). Item difficulty parameter values ranged from -2.7 to +2.7 for the 10-item condition, and for the 30-item condition, they were randomly generated based on a uniform distribution in the range of -3 to +3. Guess parameters were generated for the 0.25, 0.2, and 0.1 corresponding to easy items, medium difficulty items, and difficult items, respectively (Li et al., 2009).

Item difficulty parameter values were written in the Mplus input file as the first threshold and guessing parameter values as the second threshold (Muthén & Muthén, 1998-2021). Similar to the study of Li et al. (2009), item discrimination parameters were set as 1 for the poor and average performing classs and 2 for the good performing class. Ability parameters were obtained from the standard normal distribution (N(0,1)) and randomly generated with the runif function. In Table 1, the item parameter values generated for 10 items in the Mixture IRT models are given.

Table 1. *Item parameter values generated for the 10 items in Mixture IRT models.*

	Class1			Class2			Class1			Class2			Class3		
Item	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
1	2	-2.7	0.10	1	2.7	0.25	2	-2.7	0.10	1	-0.5	0.20	1	2.7	0.25
2	2	-2.1	0.10	1	2.1	0.25	2	-2.1	0.10	1	-0.4	0.20	1	2.1	0.25
3	2	-1.5	0.10	1	1.5	0.25	2	-1.5	0.10	1	-0.3	0.20	1	1.5	0.25
4	2	-0.9	0.10	1	0.9	0.25	2	-0.9	0.10	1	-0.2	0.20	1	0.9	0.25
5	2	-0.3	0.20	1	0.3	0.20	2	-0.3	0.20	1	-0.1	0.20	1	0.3	0.20
6	1	0.3	0.20	2	-0.3	0.20	1	0.3	0.20	1	0.1	0.20	2	-0.3	0.20
7	1	0.9	0.25	2	-0.9	0.10	1	0.9	0.25	1	0.2	0.20	2	-0.9	0.10
8	1	1.5	0.25	2	-1.5	0.10	1	1.5	0.25	1	0.3	0.20	2	-1.5	0.10
9	1	2.1	0.25	2	-2.1	0.10	1	2.1	0.25	1	0.4	0.20	2	-2.1	0.10
10	1	2.7	0.25	2	-2.7	0.10	1	2.7	0.25	1	0.5	0.20	2	-2.7	0.10

In Table 1, item parameter values generated according to all class numbers are presented for cases where the number of latent classes is two and three. For the two-class case, arranging the item difficulty parameters from easy to difficult in Class 1 means that the individuals in Class 1 produced a poorer performance when answering the items correctly, whereas arranging the item difficulty parameters from difficult to easy in Class 2 means that the individuals in Class 2 performed better when answering the items correctly. In both classes, item discrimination and guessing parameters were found to be compatible with item difficulty values. For the three-class case the item difficulty parameters in Class 2 are of medium difficulty, which means that the individuals in Class 2 produced an average performance in answering the items correctly. In all three classes, item discrimination and guessing parameters were found to be compatible with item difficulty values.

2.1.4. *Sample Size*

In previous simulation studies, sample sizes larger than 500 were selected (Lee et al., 2021; Li et al., 2009) for mixture models in simulation studies. More specifically, Li et al. (2009) reported that a sample size of 600 would be appropriate when the number of items is between 15 and 30 for the Mixture Rasch models; they also suggested that a sample size of 600 would be sufficient for a model with 1 to 4 classes for both Mixture 2PL and Mixture 3PL models for a 15-item test. Cho et al. (2013) suggested that a sample size larger than 360 can be used for the Mixture

Rasch model. Cohen and Bolt (2005) successfully applied the Mixture 3PL model with a sample size of 1000. Considering these, the sample size of the study was determined as 600 and 1000.

2.1.5. *Missing data*

Rubin (1976) classified missing data as completely at random (MCAR) and missing at random (MAR) and these missing data mechanisms have no systematic cause if they are ignored; that is, the missing data is a simple random sample of the observed data. However, if the missing pattern is missing not at random (MNAR), in this case ignoring nonignorable missing responses leads to biased parameter estimates (Little & Rubin, 1987).

In the scope of this study, MAR and MNAR data generation was based on the study of Finch (2008): for a 10-item data set, 3 most difficult items were set as target items. A total score was calculated for the remaining 7 items. Based on the total scores excluding the target items, the simulations were divided into four fractiles (0-1, 2-3, 4-5, 6-7) for each class. Four fractiles were created with four different values of the missing response probabilities on the target items. The mean of these probabilities for the fractiles was designed to be equal to the total percentages of missing responses, namely 10% and 20%. Generating missing data through this way, response patterns were formed for poor, average, and good performing simulatives based on the total scores of the items excluding the target items.

2.2. Estimation

Parameters for mixture IRT models can be estimated by Bayesian estimation with Markov chain Monte Carlo (MCMC) algorithms or maximum likelihood estimation (MLE) techniques. There are some differences in the way these two techniques are implemented. Edwards and Finch (2018) stated that the Full Information Maximum Likelihood (FIML) method produced better results in their study where they examined the parameter estimations for MAR and MNAR cases by considering the 2PL and 3PL IRT models. As the name suggests, FIML method estimates model parameters using a maximum likelihood fitting function with all the data available. Thus, individuals with missing data are included in the parameter estimation process with all the information related to them, and these are ignored for variables with missing values. In addition, FIML does not involve the assignment of missing values, thus making the use of this method less cumbersome than some of the other proposed approaches, especially those that rely on data assignment. Finally, FIML is available in most statistical software, which, in practical terms, makes it very easy to use.

2.3. Analysis

The data were analyzed with the *MplusAutomation* package, which can integrate between the Mplus program and the R program (Hallquist & Wiley, 2018). Input files to be used for 100 replications and output files obtained were also produced with the *MplusAutomation* package and analyzed. In this simulation study, the performance of Mixture IRT models was evaluated on the basis of two criteria: Item parameter recovery and classification accuracy.

2.3.1. *Item parameter recovery*

In this study, root mean square error (RMSE) values were used to assess the accuracy of item parameter estimates, calculated with the help of the following equation by using the item number, the number of classes, and the number of replications for the estimated item difficulty parameter values:

$$RMSE(\beta_i) = \sqrt{\frac{\sum_{r=1}^{R}\sum_{i=1}^{I}\sum_{g=1}^{C}\left(\hat{\beta}_{igr} - \beta_{ig}\right)^2}{RIC}}$$

In this equation, $\hat{\beta}_{igr}$ represents the estimated item difficulty parameter obtained from R replication for item i in class g, β_{ig} represents the true value of item parameter for item i in class g, R denotes the number of replications, I indicates the number of items, and C denotes the number of classes. Equation 1 was also used for the 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 with mean equating (Kolen & Brennan, 2004). The parameter estimates are exactly the same as the true value when RMSE equals zero. Lower values (e.g., <0.10) indicate better fit.

2.3.2. *Effect size*

The effect size is defined as the variance ratio describing each main effect, relationship, and error in the ANOVA design and takes a value between 0.00 and 1.00 (Cohen, 1988). Eta-square, which does not require the assumption of linearity between the variables, shows how effective the independent variable is on the dependent variable. According to Cohen (1988), 0.01 for the small effect size value; 0.06 for the medium effect size value; and 0.14 for high effect size value are recommended as lower limit values. In the presence of more than one estimator, partial eta-squared measures the proportion of the total variance explained by a given estimator, after keeping the variance explained by other estimators constant. It is recommended to use partial eta-square to determine interaction effects in multi-way or factorial ANOVA designs (Richardson, 2011; Norouzian & Plonsky, 2018). In this study, the mean RMSE values obtained from the estimated item parameters were taken as the dependent variable and the factors were also taken as independent variable. Main and interaction effects were interpreted with eta-squared values in line with the values suggested by Cohen (1988).

2.3.3. *Classification accuracy*

Within the data sets produced for classification accuracy, there is a posterior probability for each person in each latent class based on person's response pattern. Each person in the latent class was assigned to a latent class according to their highest posterior probability values, saved in the Mplus output and these values were extracted with the MplusAutomation package. For a data set with 1000 examinees, classification accuracy value was calculated as 0.92, which means there is a matched assignment for 920 of the 1000 cases.

2.3.4. *Label switching*

Since there is no information about the number and nature of estimated classes in mixture IRT models, sometimes the parameters estimated for Class 1 can be labeled as Class 2. In such cases, the problem of label switching can be overcome by taking the estimated item parameter values as starting values in Mplus syntax (Kutscher et al. 2019).

3. RESULTS

3.1. Item Parameter Recovery Results

3.1.1. *Item difficulty parameter*

The mean RMSE values of the estimated item difficulty parameters for the mixture models are presented in Table 2. The codes in this table for simulation conditions are designed to represent the combination of factors for a given situation. To specify the simulation conditions, codes with 10-13 digits were created. The first two characters of the codes denote class number (2C, 3C); the following three characters refer to missing data percentage (10P, 20P); the next grouping indicates sample size (600, 1000), and the last two characters represent the number of items (10,30). For example, in the 2C10P60010 codes the number of classes is denoted by 2C, the percentage of missing data by 10P, the sample size by 600, and the number of items by 10.

Table 2. *The mean RMSE values of the estimated item difficulty parameters for the Mixture models.*

	Mixture Rasch			Mixture 2PLM			Mixture 3PLM		
Conditions	COMP	MAR	MNAR	COMP	MAR	MNAR	COMP	MAR	MNAR
2C10P60010	0.045	0.052	0.075	0.041	0.065	0.089	0.367	0.560	0.373
2C10P60030	0.025	0.026	0.054	0.025	0.039	0.077	0.112	0.225	0.303
2C10P100010	0.035	0.043	0.067	0.032	0.038	0.070	0.219	0.265	0.263
2C10P100030	0.021	0.022	0.050	0.021	0.022	0.052	0.090	0.168	0.352
2C20P60010	0.046	0.083	0.082	0.057	0.121	0.241	0.348	0.654	0.592
2C20P60030	0.025	0.028	0.251	0.037	0.061	0.070	0.130	0.213	0.222
2C20P100010	0.035	0.071	0.540	0.034	0.233	0.177	0.219	0.586	0.520
2C20P100030	0.023	0.024	0.188	0.023	0.024	0.919	0.078	0.124	0.217
3C10P60010	0.139	1.160	0.236	1.551	1.775	1.950	1.386	1.898	2.831
3C10P60030	0.105	0.435	0.170	0.070	0.032	0.064	0.210	0.221	0.257
3C10P100010	0.102	0.206	0.125	0.973	1.256	1.778	1.075	1.704	2.636
3C10P100030	0.082	0.069	0.036	0.051	0.096	0.075	0.160	0.196	0.195
3C20P60010	0.148	1.350	1.690	1.761	1.994	2.570	1.160	4.471	2.357
3C20P60030	0.081	0.397	0.191	0.279	1.436	1.709	0.246	0.292	0.283
3C20P100010	0.090	0.884	0.383	1.641	1.832	2.237	2.652	2.673	2.341
3C20P100030	0.024	0.058	0.088	0.226	0.439	0.113	0.141	0.253	0.212

Table 2 shows that the mean RMSE values of the item difficulty parameters obtained for the Mixture Rasch model decreased as the number of items and the number of classes increased. As can be seen in Table 2, in the complete data, the mean RMSE values decreased as the number of items and sample size increased, and the mean RMSE values increased as the number of classes and the percentage of missing data increased. In MAR and MNAR data conditions, the mean RMSE values generally decreased as the number of items and sample size increased, and the mean RMSE values generally increased as the number of classes and the percentage of missing data increased. It can also be seen that item difficulty parameter values had the highest mean RMSE values in complete, MAR, and MNAR data with 3 class, 20% missing data percentage, 600 sample size, and 10 item (3C20P60010) condition. The lowest mean RMSE value was observed in complete data, 2 class, 10% missing data, 1000 sample size, and 30 item (2C10P100030) condition.

In Table 2, it can be seen that the mean RMSE values of the item difficulty parameters obtained for the Mixture 2PL model were higher than the mean RMSE values of the item difficulty parameters obtained for the Mixture Rasch model. Also, the mean RMSE values decreased as the number of items and sample size increased, and the mean RMSE values increased as the number of classes and the percentage of missing data increased. When the mean RMSE values were examined according to the missing data types, higher RMSE values were obtained for the MNAR condition. The item difficulty parameter values for the mixture 2PL model were obtained with the highest RMSE values, while the MAR and MNAR data with 3 class, 20% missing data percentage, 600 sample size, and 10 item (3C20P60010) condition. The lowest mean RMSE value was observed in the complete data with 2 class, 10% missing data, 1000 sample size, and 30 item (2C10P100030) condition.

As shown in Table 2, the mean RMSE values of the item difficulty parameters obtained for the Mixture 3PL model were higher than those for the Mixture 2PL model. Also, the mean RMSE values increased as the complexity of the model increased (i.e from Rasch to 3PL model). The mean RMSE values decreased as the number of items and sample size increased, and the mean RMSE values increased as the number of classes and the percentage of missing data increased.

When the mean RMSE values were examined according to the missing data types, higher RMSE values were obtained for the MAR data condition. Table 2 shows that the highest mean RMSE value of the item difficulty parameter values for the mixture 3PL model was MAR data, with 3 class, 20% missing data percentage, 10 item, and a sample size of 600 (3C20P60010) condition and also the lowest RMSE value was seen with complete data, 2 class, 10% missing data, 30 item, and a sample size of 1000 (2C10P100030) condition.

3.1.2. *Item discrimination parameter*

The mean RMSE values of item discrimination parameter values for the Mixture 2PL and 3PL model are given in Table 3:

Table 3. *The mean RMSE values of the estimated item discrimination parameters for the Mixture 2PL and 3PL models.*

	Mixture 2PLM			Mixture 3PLM		
Conditions	COMP	MAR	MNAR	COMP	MAR	MNAR
2C10P60010	0.165	0.212	0.497	0.331	0.645	0.634
2C10P60030	0.179	0.310	0.277	0.161	0.234	0.171
2C10P100010	0.057	0.061	0.063	0.264	0.320	0.216
2C10P100030	0.024	0.038	0.052	0.094	0.115	0.120
2C20P60010	0.234	0.261	0.563	0.371	0.654	0.662
2C20P60030	0.259	0.256	0.523	0.192	0.229	0.262
2C20P100010	0.183	0.197	0.208	0.337	0.405	0.417
2C20P100030	0.032	0.041	0.055	0.101	0.151	0.176
3C10P60010	0.843	0.937	1.222	1.140	1.293	1.337
3C10P60030	0.725	0.873	0.916	0.776	0.821	0.857
3C10P100010	0.866	0.910	1.469	0.988	1.113	1.228
3C10P100030	0.675	0.784	0.833	0.581	0.696	0.705
3C20P60010	0.975	1.277	1.366	1.262	1.463	1.472
3C20P60030	0.837	0.927	0.982	0.920	0.943	0.952
3C20P100010	0.922	0.981	1.032	1.023	1.242	1.281
3C20P100030	0.786	0.854	0.967	0.723	0.817	0.832

As shown in Table 3, the mean RMSE values of the item discrimination parameter estimations for the complete data condition were lower, slightly higher for the MAR condition, and at the highest for the MNAR condition. The lowest RMSE values were obtained for complete, MAR and MNAR data with for 2 class, 10% missing data, 1000 sample size, and 30 item (2C10P100030) condition, while the highest RMSE value was obtained for the MNAR data with 3 class, 10% missing data percentage, 1000 sample size, and 10 item (3C10P100010) condition. It seems to be consistent with the conditions where the highest RMSE values were obtained for item discrimination parameter estimations and the highest mean RMSE values for item difficulty parameter estimations. For the mixture 3PL model, the mean RMSE values were lower for the complete data case of item discrimination parameter estimations, but higher for the MAR and MNAR conditions. The lowest RMSE values were obtained for complete, MAR, and MNAR data with 2 class, 10% missing data, 1000 sample size, and 30 item (2C10P100030) condition, while the highest RMSE value was obtained for MAR and MNAR data with data with 3 class, 20% missing data percentage, 600 sample size, and 10 item (3C10P100010) condition. It can be stated that these results and the conditions in which the highest RMSE values were obtained for item discrimination and item difficulty parameter estimation values in the Mixture 2PL model were similar.

3.1.3. *Guessing parameter*

Table 4 provides the mean RMSE values obtained for the guessing parameter values for mixture 3PL model.

Table 4. *The Mean RMSE values of the estimated guessing parameters.*

Conditions	COMP	MAR	MNAR
2C10P60010	0.076	0.076	0.076
2C10P60030	0.046	0.046	0.046
2C10P100010	0.077	0.077	0.078
2C10P100030	0.046	0.046	0.046
2C20P60010	0.076	0.076	0.077
2C20P60030	0.046	0.046	0.046
2C20P100010	0.078	0.078	0.079
2C20P100030	0.046	0.046	0.046
3C10P60010	0.059	0.059	0.060
3C10P60030	0.038	0.038	0.039
3C10P100010	0.061	0.058	0.061
3C10P100030	0.038	0.039	0.039
3C20P60010	0.059	0.059	0.061
3C20P60030	0.039	0.038	0.039
3C20P100010	0.061	0.061	0.061
3C20P100030	0.039	0.039	0.039

It can be seen in Table 4 that when the number of items for the guessing parameter increased, the mean RMSE values decreased as well. Also, the mean RMSE values for guessing parameters had lower values than the mean RMSE values obtained for item difficulty and discrimination parameters. The reason for this could be that when the guessing parameter values are between zero and one, item discrimination and difficulty parameter values can take larger absolute values.

3.2. A Linear Model Analysis of Simulation Results

Effects of each condition were evaluated using a factorial ANOVA for the RMSE values. The results related to partial eta-squared, degree of freedom (*df*), sum of squares (SS), mean square (MS), and F-values from the factorial ANOVA are presented in the following sections.

3.2.1. *ANOVA Results for item difficulty parameter*

In Table 5, main effects, two-way and three-way interactions for each factor are shown for item difficulty parameter. As can be seen in Table 5, all factors had a significant effect on item parameter estimation. According to partial eta-squared values, number of items (i), number of classes (C), and model (M) were the most influential factors on RMSE for item difficulty parameter. Missing data type and missing data percentage had also a large effect on the results. The least influential factor was the sample size (N).

The interaction effects between factors shown in Table 5 indicate that type and class (txC), type and percentage (txP), item and class (ixC), item and model (ixM), sample and class (NxC), class and model (CxM), and percentage and model (PxC) affected the RMSE values. Based on partial eta-squared values, it can be seen that two-way interactions had a large effect on the results. Also, significant three-way interactions are given in Table 5 and it can be seen that type, item and class (txixC), type, class and model (txCXM), item, class and model (ixCxM), and

sample, class and model (NxCxM) had significant interaction effects. These results suggest that interactions of factors may affect model parameter estimates.

Table 5. *ANOVA results for main effects and interaction effects of simulation conditions for item difficulty parameter.*

Factor	η_p^2	*df*	Sum of Squares	Mean Square	F
t	0.832	2	3.881	1.940	54.318*
i	0.922	1	9.313	9.313	260.685*
N	0.540	1	0.922	0.922	25.798*
C	0.951	1	15.340	15.340	429.413*
P	0.702	1	1.854	1.854	51.902*
M	0.921	2	9.126	4.563	127.724*
txi	0.085	2	0.073	0.037	1.024
txN	0.039	2	0.032	0.016	0.444
txC	0.663	2	1.543	0.771	21.593*
txP	0.426	2	0.582	0.291	8.151*
txM	0.324	4	0.377	0.094	2.635
ixN	0.109	1	0.096	0.096	2.690
ixC	0.767	1	2.581	2.581	72.259*
ixP	0.005	1	0.004	0.004	0.106
ixM	0.489	2	0.752	0.376	10.531*
NxC	0.173	1	0.164	0.164	4.591*
NXP	0.016	1	0.013	0.013	0.360
NXM	0.077	2	0.066	0.033	0.920
CXP	0.007	1	0.006	0.006	0.165
CXM	0.687	2	1.723	0.862	24.119*
PXM	0.459	2	0.667	0.333	9.334*
txixC	0.440	2	0.617	0.309	8.637*
txCxM	0.548	4	0.954	0.239	6.676*
ixCXM	0.439	2	0.615	0.307	8.605*
NXCXM	0.268	2	0.288	0.144	4.025*
Error		22	0.786	0.036	

Note. *t* = missing data type, i = number of items, N = sample, C =number of classes, P=missing data percentage, M = model.
*$p<.05$.

3.2.2. *ANOVA Results for item discrimination parameter*

In Table 6, main effects, two-way and three-way interactions for each factor are shown for item disrimination parameter. As can be seen in Table 6, according to partial eta-squared values, all factors had a large effect size values on the results. Mean RMSE values for item discrimination parameter were also significantly affected by two-way interactions including type and item (txi), type and class (txC), item and class (ixC), type and model (txM), item and class (ixC), item and model (ixM), and sample and class (PxC). Based on partial eta-squared values, these interactions had a large effect on the results. Also, three-way inreactions type, sample and class (txNxC), and sample, class and model (NxCXM) affected mean RMSE values for item discrimination parameter. These results suggest that interactions of factors may affect model parameter estimates.

Table 6. *ANOVA results for main effects and interaction effects of simulation conditions for item discrimination parameter.*

Factor	η_p^2	*df*	Sum of Squares	Mean Square	F
t	0.905	2	0.445	0.223	57.014*
i	0.969	1	1.442	1.442	369.207*
N	0.923	1	0.565	0.565	144.609*
C	0.996	1	13.065	13.065	3345.500*
P	0.760	1	0.149	0.149	38.052*
M	0.782	1	0.168	0.168	43.128*
txi	0.593	2	0.068	0.034	8.756*
txN	0.379	2	0.029	0.014	3.658
txC	0.428	2	0.035	0.018	4.483*
txP	0.012	2	0.001	0.000	0.072
txM	0.558	2	0.059	0.030	7.570*
ixN	0.099	1	0.005	0.005	1.318
ixC	0.777	1	0.163	0.163	41.723*
ixP	0.007	1	0.000	0.000	0.084
ixM	0.856	1	0.279	0.279	71.379*
NxC	0.404	1	0.032	0.032	8.141*
NXP	0.123	1	0.007	0.007	1.686
NXM	0.021	1	0.001	0.001	0.258
CXP	0.103	1	0.005	0.005	1.379
CXM	0.185	1	0.011	0.011	2.716
PXM	0.022	1	0.001	0.001	0.268
txNxC	0.489	2	0.045	0.022	5.738*
NxCXM	0.495	1	0.046	0.046	11.774*
Error		12	0.047	0.004	

Note. t = missing data type, i = number of items, N = sample, C =number of classes, P=missing data percentage, M = model.
*p<.05.

3.2.3. *ANOVA Results for guessing parameter*

In Table 7 for each factor, main effect, two-way, and three-way interactions are shown for guessing parameter.

Table 7. *ANOVA results for main effects and interaction effects of simulation conditions for guessing parameter.*

Factor	η_p^2	*df*	Sum of Squares	Mean Square	F
t	0.951	2	0.000	0.000	19.316
i	1.000	1	0.008	0.008	42941.408*
N	0.64	1	0.000	0.000	54.039*
C	1.000	1	0.002	0.002	9336.320*
P	0.769	1	0.000	0.000	6.671
ixC	0.999	1	0.000	0.000	1547.526*
Error		2	0.000	0.000	

Note. *t* = missing data type, i = number of items, N = sample, C =number of classes, P=missing data percentage.
*p<.05.

When Table 7 is examined, according to partial eta-squared values, it can be seen that especially item and class factors had a large effect on the results, but main effects of missing data type and missing data percentage were found to have no significant effects. Mean RMSE values for item guessing parameter were also significantly affected by interaction between item and class factors.

3.3. Classification Accuracy Results

Table 8 shows the classification rates for mixture IRT models.

Table 8. *The Classification rates for the Mixture Models.*

	Mixture Rasch			Mixture 2PLM			Mixture 3PLM		
Conditions	COMP	MAR	MNAR	COMP	MAR	MNAR	COMP	MAR	MNAR
2C10P60010	98.62	85.03	81.06	98.36	86.85	72.54	98.51	83.71	77.93
2C10P60030	93.01	87.48	83.07	98.71	87.35	75.44	98.77	88.79	79.98
2C10P100010	98.62	85.26	84.58	98.34	86.71	72.79	96.90	83.87	77.92
2C10P100030	99.02	87.81	82.93	99.02	87.41	76.06	99.58	89.30	81.33
2C20P60010	98.60	79.98	72.03	98.26	72.18	65.02	95.66	81.20	68.47
2C20P60030	89.02	76.59	72.93	97.03	77.01	72.69	97.01	72.30	73.18
2C20P100010	98.65	75.61	75.70	98.33	72.97	64.90	96.22	67.25	68.26
2C20P100030	88.01	76.56	62.44	86.02	77.08	72.90	99.60	75.88	70.46
3C10P60010	85.26	64.49	64.87	83.70	75.34	68.44	84.37	61.45	61.54
3C10P60030	83.32	74.43	75.30	83.92	83.47	79.83	83.32	70.96	68.79
3C10P100010	88.58	67.72	71.61	86.51	76.82	72.01	86.76	62.20	65.25
3C10P100030	84.92	75.85	75.54	84.92	84.18	80.15	84.92	74.38	71.22
3C20P60010	77.66	61.75	59.25	83.33	71.42	67.10	80.21	62.64	66.75
3C20P60030	83.70	63.81	61.86	93.70	72.14	76.08	81.80	62.28	68.33
3C20P100010	78.60	65.20	60.83	91.66	74.86	68.05	84.65	66.06	69.67
3C20P100030	85.26	64.44	62.54	93.24	71.75	77.83	82.53	63.53	70.46

As can be seen in Table 8, higher classification accuracy percentages were obtained for the compelete data case in the Mixture Rasch model. In the complete data condition, the highest percentage of classification accuracy was achieved for 2 class with 10% of missing data, 30 item, and a sample size of 1000 (99.02), while the lowest percentage of classification accuracy was achieved for 3 class with 20% missing data, 10 item, and a sample size of 600 (77.67). According to the missing data type, lower classification accuracy percentages were obtained in MAR and MNAR pattern conditions. In the MAR pattern condition, the highest percentage of classification accuracy was achieved for 2 class, 10% missing data, 30 item, and a sample size of 1000 (87.81), while the lowest percentage of classification accuracy was obtained 3 class, 20% missing data, 10 item, and a sample size of 600 (61.75) condition. In the MNAR pattern condition, the highest percentage of classification accuracy was reached for 2 class, 10% missing data, 10 item, and a sample size of 1000 (84.58), while the lowest percentage of classification accuracy was found 3 classes, 20% missing data, 10 item, and a sample size of 600 (59.25) condition.

For the mixture 2PL model condition, higher percentages of classification accuracy were obtained for the complete data case. In the complete data condition, the highest percentage of classification accuracy was achieved in combinations of 2 class, 10% of missing data, 30 item, and a sample size of 1000 item condition (99.02), while the lowest percentage of classification accuracy was found for 3 class, 20% missing data, 10 item, and a sample size of 600 (83.32). According to the missing data type, lower classification accuracy percentages were obtained in

the MNAR pattern condition. In the MAR pattern condition, the highest percentage of classification accuracy was achieved for 2 class, 10% of missing data, 30 item, and a sample size of 1000 (87.41) while the lowest percentage of classification accuracy was achieved for 3 class, 20% missing data, 10 item, and a sample size of 600 (71.42) condition. In the MNAR pattern condition, the highest percentage of classification accuracy was achieved for 3 class, 10% of missing data, 30 item, and a sample size of 1000 (80.15), while the lowest percentage of classification accuracy was found for 3 class, 20% missing data, 10 item, and a sample size of 600 (64.90) condition.

For the mixture 3PL model condition, higher percentages of classification accuracy were obtained for the complete data case as well. In the complete data condition, the highest percentage of classification accuracy was achieved for the combinations of 10% (99.58) and 20% (99.60) missing data percentages of 2 class with 30 item and a sample size of 1000 condition. The lowest percentage of classification accuracy was achieved for 3 class, 20% missing data, 10 item, and a sample size of 600 (80.213) condition. According to the missing data type, lower classification accuracy percentages were obtained under MAR and MNAR missing data pattern conditions. The highest percentage of classification accuracy obtained for the MAR and MNAR missing data pattern was 2 class, 10% of missing data was in the condition of 30 item and a sample size of 1000, and the lowest percentage of classification accuracy was in 3 class, 10% missing data percentage, 10 item, and a sample size of 600 condition.

4. DISCUSSION and CONCLUSION

Although mixture IRT models have been found to be useful in the fields of pscyhology, education and medicine, little research has been reported on the effects of sample size, number of items, number of latent classes, missing data, factors on model parameter estimates, and classification accuracy. In this research, a simulation study was conducted to examine the effects of estimation model, the number of items, sample size, the number of latent classes, missing data type, the percentage of missing data conditions on item parameter recovery, and classification accuracy for three mixture IRT models. The mean RMSE values were examined for parameter recovery. Furthermore, the main effects and interaction effects of the factors were examined. In addition, classification accuracy percentages were obtained by comparing the estimated latent class memberships with the true class memberships.

The findings indicate that, in the estimation of item difficulty and discrimination parameters for mixture IRT models, lower mean RMSE values were obtained as the sample size and number of items increased; on the other hand, the mean RMSE value increased as the number of classes increased. In the estimation of the guessing parameter, it was seen that the mean RMSE value decreases as the sample size, number of items and classes increase. These results match the ones observed in other studies. Previous studies investigating the effect of sample size, number of items, number of classes on parameter recovery for Mixture IRT models on item difficulty, and item discrimination parameter estimation (Alexeev et al., 2011; Cho et al., 2013; Finch & French, 2012; Li et al., 2009; Preinerstorfer & Formann , 2012; Sen et al., 2016) found that the mean RMSE value decreased as the sample size and number of items increased, and the mean RMSE values increased as the number of classes increased. In the estimation of the guessing parameter, it was observed that the mean RMSE values decreased as the sample size, number of items, and number of classes increased (Finch & French, 2012; Sen et al., 2016). It can be said that the results of this study are consistent with those in the related literature. It has been suggested that when the number of classes increases, it is natural for the error values to increase due to the decrease in the number of individuals in the classes (Finch & French, 2012).

In the item difficulty and item discrimination parameter estimations for the mixture models, lower mean RMSE values were obtained for the complete data cases, and higher mean RMSE

values were obtained for the MAR and MNAR cases of the missing data type. In the cases where the percentage of missing data was 20%, higher mean RMSE values were achieved. Similar results were found in a study in the literature in which mixture Rasch and mixture 2PL model and missing data type and percentage conditions were discussed (Lee, 2012). Obtaining mean RMSE values close to each other for the guessing parameter according to the missing data types corroborate the findings of Finch (2008), where the mean RMSE values of MAR and MNAR conditions were found to be low and very close to each other in the estimation of guessing parameter for IRT models. Since the missing data generation mechanism was produced as in Finch (2008), and the missing data were analyzed by the FIML method without assigning missing data, it seems natural that the mean RMSE values for the guessing parameter are close to each other.

In the recovery of the item parameters, it was observed that the mean RMSE values obtained for the Mixture 3PL model were higher than the mean RMSE values obtained for the Mixture 2PL model, and that the mean RMSE values obtained for the Mixture 2PL model were higher than the mean RMSE values obtained for the Mixture Rasch model. In the estimation of item parameters, a pattern of RMSE values appears to increase as the complexity of the model increases. Therefore, it can be said that the item parameters obtained for the Mixture Rasch model have fewer errors than those of the Mixture 2PL and 3PL models. These results are in agreement with the studies that obtained lower RMSE values for the two-class Mixture Rasch model (Cho et al., 2013; Sen, 2014).

In addition, when the main effects and interaction effects of the factors were examined, significant and high effect size values were obtained for the main effects of all factors considered in the estimation of item difficulty and discrimination parameters; however, for guessing parameter, it was obtained only for item, class, and model factors. These results suggest that interactions of factors may affect model parameter estimates and factors with high effect size values are important factors.

When the classification accuracy percentages were examined, higher classification accuracy percentages were obtained for the complete data case in all the Mixture IRT models. For all the mixture IRT models, in the complete data and MAR data condition, the highest percentage of classification accuracy was obtained in the combinations of 2 class, 10% missing data, 30 item, and a sample size of 1000, while the lowest classification accuracy was reached for the 3 class, 20% missing data percentage, 10 item, and a sample size of 600 condition. It was observed that lower classification accuracy was obtained for all the models in MAR and MNAR conditions.

5. SUGGESTION AND LIMITATIONS

The values used in the generation of item difficulty, item discrimination, and guessing parameters in this specific study are limited to the values used in the study of Li et al. (2009). In further studies researchers can change the item parameter generating values using different distributions. In addition, in this research, it is assumed that the ability parameter is randomly obtained from the standard normal distribution; using different ability distributions, researchers can examine the accuracy of recovery of item paremeters. In our simulation study 100 replications were performed for each condition; researchers can interpret the analysis results by changing the number of replications. In this study, the analysis of missing data was carried out using FIML method without using missing data assignment methods; by using missing data assignment methods, researchers can examine the effects of these methods in Mixture IRT models. In addition, the MLR estimation method was used for the estimation of the parameters; researchers can use different methods such as Bayesian and these methods can be compared. The results of this study are based on dichotomously scored items; researchers can perform Mixture IRT models analyses with polytomous scored items.

Acknowledgments

This paper was produced from the first author's doctoral dissertation prepared under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Fatima Munevver Saatcioglu: Investigation, Resources, Visualization, Software, Analysis, and Writing-original draft. **Hakan Yavuz Atar**: Methodology, Supervision, and Critical Review.

Orcid

Fatima Munevver Saatcioglu https://orcid.org/0000-0003-4797-207X
Hakan Yavuz Atar https://orcid.org/0000-0001-5372-1926

REFERENCES

Alexeev, N., Templin, J., & Cohen, A.S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, *48*, 313–332.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Academic.

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

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*, 278–306. https://doi.org/10.1080/00949655.2011.603090

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. https://doi.org/10.1080/08957347.2012.714682

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

Choi Y.J., & Cohen, A.S. (2020). Comparison of scale identification methods in Mixture IRT models. *Journal of Modern Applied Statistical Methods*, *18*(1), eP2971. https://doi.org/10.22237/jmasm/1556669700

Collins, L.M., & Lanza, S.T. (2010). *Latent class and latent transition analysis*. John Wiley & Sons.

De Ayala, R.J., Plake, B.S. & Impara, J.C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement,38*, 213–234. https://doi.org/10.1111/j.1745-3984.2001.tb01124.x

De Ayala, R.J. & Santiago, S.Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*, *60*, 25-40. https://doi.org/10.1016/j.jsp.2016.01.002

Edwards, J.M., & Finch, W.H. (2018). Recursive partitioning methods for data imputation in the context of item response theory: *A Monte Carlo simulation. Psicológica, 39(1)*, 88-117. https://doi.org/10.2478/psicolj-2018-0005

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225-245. https://doi.org/10.1111/j.1745-3984.2008.00062.x

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.

Hallquist, M.N., & Wiley, J.F. (2018). MplusAutomation: An R package for facilitating large scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 25(4)*, 621-638. https://doi.org/10.1080/10705511.2017.1402334

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Sage.

Hohensinn, C., & Kubinger, K.D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*, 732-746. https://doi.org/10.1177/0013164410390032

Jilke, S., Meuleman, B., & Van de Walle, S. (2015). We need to compare, but how? Measurement equivalence in comparative public administration. *Public Administration Review*, *75*(1), 36–48. https://doi.org/10.1111/puar.12318

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.

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. https://doi.org/10.3389/fpsyg.2019.02494

Lee, S. (2012). *The Impact of Missıng Data on The Dichotomous Mixture IRT Models* [Unpublished Doctoral Dissertation]. The University of Georgia.

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

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

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. Wiley.

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*, 611-631. https://doi.org/10.1177/0146621607312613

Muthén, L.K. & Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*, *34*, 257–271. https://doi.org/10.1177/0267658316684904

Oliveri, M.E., Ercikan, K., Zumbo, B.D., & Lawless, R. (2014). Uncovering substantive patterns in student responses in international large-scale assessments-Comparing a latent class to a manifest DIF approach. *International Journal of Testing*, *14*(3), 265-287. https://doi.org/10.1080/15305058.2014.891223

Park, Y.S., Lee, Y.-S., & Xing, K. (2016). Investigating the impact of item parameter drift for item response theory models with mixture distributions. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.00255

Pohl, S., Grafe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74(3)*, 423–452. https://doi.org/10.1177/0013164413504926

Preinerstorfer, D., & Formann, A.K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, *65*, 251–262. https://doi.org/10.1111/j.2044-8317.2011.02020.x

R Core Team (2020). *R: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria. https://www.R-project.org/

Richardson, J.T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review 6*, 135-47. https://doi.org/10.1016/j.edurev.2010.12.001

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

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost, & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). Waxmann.

Sen, S. (2014). *Robustness of mixture IRT models to violations of latent normality*. [Unpublished Doctoral Dissertation]. The University of Georgia.

Sen, S., Choen, 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

Sen, S., & Cohen, A.S. (2019). Applications of mixture IRT models: A literature review, *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177-191. https://doi.org/10.1080/15366367.2019.1583506

Zhang, D., Orrill, C., & Campbell, T. (2015). Using the mixture Rasch model to explore knowledge resources students invoke in mathematics and science assessments. *School Science and Mathematics*, *115*(7), 356-365. https://doi.org/10.1111/ssm.12135