Participatory Educational Research (PER)
Vol. 7(3), pp. 180-191, December 2020
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.20.41.7.3

A Study on the Identification of Latent Classes Using Mixture Item Response Theory Models: TIMSS 2015 Case

Fatıma Münevver Saatçioğlu*
Ankara Yıldırım Beyazıt University, Ankara, Turkey
ORCID: 0000-0003-4797-207X

Hakan Yavuz Atar
Gazi Faculty of Education, Gazi University, Ankara Turkey
ORCID: 0000-0001-5372-1926

Article history

Received:
26.04.2020

Received in revised form:
11.06.2020

Accepted:
24.06.2020

Key words:
Mixture item response theory;
latent class;
heterogeneity;
TIMSS 2015;
validity

This study examined the existence of latent classes in TIMSS 2015 data from three countries, Singapure, Turkey and South Africa, were analyzed using Mixture Item Response Theory (MixIRT) models (Rasch, 1PL, 2PL and 3PL) on 18 multiple-choice items in the science subtest. Based on the findings, it was concluded that the data obtained from TIMSS 2015 8th grade science subtest have a heterogeneous structure consisting of two latent classes. When the item difficulty parameters in two classes were examined for Singapore, it was determined that the items were considerably easy for the students in Class 1 and the items were easy for the students in Class 2. When the item difficulty parameters in two classes were examined for Turkey, it was found that the items were easy for the students in Class 1 and the items were difficult for the students in Class 2. When the item difficulty parameters in two classes were examined for South Africa, it was ascertained that the items were a bit easy for the students in Class 1 and the items were considerably difficult for the students in Class 2. The findings were discussed in the context of the assumption of parameter invariance and test validity.

Introduction

Accurate understanding and analysis of data in education and related fields are important to obtain reliable and valid measurements and evaluations. In particular results from international large-scale assessments guide the process of education to be more efficient and allow the academic achievements of student groups in one country to be compared with those in other countries (Cook, 2006). A method based on student samples from all participating countries and calibrations of the Item Response Theory (IRT) is implemented to ensure comparability of scores in international large-scale assessments. This method ensures that each participating country contributes an equal amount to the calibration of item parameters (Oliveri & von Davier, 2011).

* Correspondency: fmsaatcioglu@ybu.edu.tr, Phone: +90 312 906 13 71

IRT models are used to accurately determine the success of students in international large-scale assessments (Martin, Mullis, & Hooper, 2016; Yamamoto & Kulick, 2000). Using IRT models, the relationship between an examinee's ability (latent variable) and the probability of the examinee responding correctly to any item is modeled (Harris, 1989). Three different IRT models are used in TIMSS assessment based on item type and scoring method. A three-parameter logistics model is used for multiple-choice items and a two-parameter logistics model for the constructed-response items that were scored as dichotomous. A generalized partial credit model is used for polytomous scored constructed-response items (Martin et al., 2016).

Although Item Response Theory models have many advantages, they have parsimonious assumptions such as unidimensionality, parameters invariance, and local independence (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). To gather accurate evidence regarding the validity of the model used in the analysis, its assumptions must be met and there must be no biased items (Kreiner & Christensen, 2007). The advantages of the IRT models rely on the validity of the model which requires its assumptions to be met. However, these assumptions are quite difficult to meet in many types of research (von Davier, Rost, & Carstensen, 2007).

The parameter invariance assumption of Item Response Theory means that the estimated item parameter values do not change over different groups (Hambleton et al., 1991; Embretson & Reise, 2000; DeMars, 2010). However, in some cases, different groups can be formed due to the response strategies and techniques that individuals use to respond to the items correctly, and these groups are defined as latent classes (Embretson, 2007; Glück & Spiel, 2007). In other words, differences among individuals in terms of different problem-solving techniques, being familiar with item contents or having different educational backgrounds, etc. can lead to the formation of different latent classes (Mislevy & Huang, 2007; Rijkes & Kelderman, 2006). The presence of many latent classes in data obtained from tests means that the measured psychological construct varies among different groups, thus threatening test validity (Kreiner & Christensen, 2007; Messick 1994; von Davier & Yamamoto, 2007; Toker, 2016). This is because it is reported that assumptions such as unidimensionality, local independence, parameter invariance, and monotonicity and the absence of items with differential item functioning (DIF) in the test can be considered as requirements of construct validity in standard IRT models (Kreiner & Christensen, 2007).

Several models such as multidimensional IRT models (Reckase, 2009), multiple group IRT models (Bock & Zimowski 1997) and Mixture IRT models (de Ayala & Santiago, 2017; Rost 1990; Mislevy & Verhelst 1990) have been developed in case assumptions of standard IRT models are violated or cannot be met. Unlike IRT models, Mixture IRT models do not require parameter invariance assumption and they allow item parameters to vary among latent classes (de Ayala & Santiago, 2017; von Davier & Yamamoto, 2007). The variation of item parameters between latent classes indicates the existence of homogeneous subgroups (Rost, 1990; de Ayala & Santiago, 2017). Analyzing heterogeneous datasets consisting of homogeneous subgroups using standard IRT models can cause misinterpretation of the results (DeMars, 2010; Finch & French, 2012).

In Mixture IRT models, item parameters and individuals' ability distributions (mean and variance) can vary in different latent classes (Rost, 1990; de Ayala & Santiago, 2017). Changes of item and ability parameters reveal that some characteristics of individuals in different classes such as strategies employed by them and familiarity with question types vary (Kreiner & Christensen, 2007; von Davier & Yamamoto, 2007). As a result, the inclusion of individuals in

different latent classes based on their abilities enables researchers to obtain more reliable and valid information about item and group traits (de Ayala & Santiago, 2017). That is because using a single and the same parameter estimation for all groups despite latent classes consisting of individuals with different ability levels causes loss of information. Furthermore, it is possible to obtain more information by simultaneously modeling both continuous (ability parameter) and categorical (latent class) data using the Mixture IRT approach (de Ayala & Santiago, 2017).

The use of Mixture IRT applications is increasing day by day in educational and psychological assessment studies. These include studies in different problem-solving strategies (Mislevy & Verhelst, 1990; Rijkes & Kelderman, 2006; Rost & von Davier, 1993), studies in DIF (Maij-de Meij, Kellerman, & van der Flier, 2010; Samuel, 2005), studies in test speededness (Bolt, Cohen & Wollack, 2002; Meyer, 2008), studies in personality questionnaires (Hong, 2007; Meiser & Machunsky, 2008; Rost, Carstensen & von Davier, 1997), and measurement invariance studies (Eid & Rauber, 2000).

Analysis of the relevant literature reveals that there are many studies examining the existence of latent classes in international large-scale test data (Choi, Alexeev, &Cohen, 2015; Liu, Liu, & Li, 2018; Oliveri, Ercikan, Zumbo, & Lawless, 2014; Oliveri & von Davier, 2011; Park, Lee, & Xing, 2016; Sen, Cohen & Kim, 2016; Toker, 2016; Zhang, Orrill, & Campbell, 2015). Three latent classes were identified in a study examining the heterogeneity in response patterns of fourth-grade students from Taiwan, Hong Kong, Qatar and Kuwait who participated in PIRLS 2006 (Oliveri et al., 2014). Choi et al. (2015) identified two latent classes using a three-parameter Mixture IRT model in an analysis that they conducted with 11 multiple choice and 15 open-ended (binary scored) items on the 4th grade mathematics data for seven countries that participated in TIMMS-2007 (Austria, Australia, El Salvador, Hong Kong, Qatar, Singapore, Slovakia). Zhang et al. (2015) conducted three separate analyses on the 15 items in the science subtest, 16 items in the mathematics subtest and the total including the science subtest, mathematics subtest and the combination of these two tests in PISA 2009 for Chinese data to gather information about the classification of students in the domains of science and mathematics. They concluded that the data obtained from the Chinese students fitted the two-class Mixture Rasch model best in each subtest and in cases where those subtests were employed together. Sen et al. (2016), on the other hand, concluded that the data obtained from the South Korean students who achieved the highest success in the 8th grade mathematics subtest in TIMMS 2011 fitted the two-class Mixture Rasch model best.

This study is important as it examines the existence of latent classes in TIMSS 2015 data, interprets model outputs under the Mixture IRT model that fits the data best, and the validity of the TIMSS assessment. Using unidimensional standard IRT models in large-scale tests such as TIMSS and PISA causes latent classes to be ignored, and therefore the parameter invariance assumption is violated (von Davier, Rost, & Carstensen, 2007). In this situation biased results can be obtained in item parameter calibrations (DeMars & Lau, 2011). Furthermore, although it is emphasized that the parameter invariance assumption is necessary for cross-cultural comparisons (Hambleton & Rogers, 1989; Meredith, 1993; Millsap, 2011), testing of assumptions for data obtained from international assessments can be neglected (Park et al., 2016).

This study investigates the heterogeneity of data from Singapore, Turkey and South Africa countries which achieved high, medium and low levels of success in TIMSS 2015 respectively. For this reason, Mixture IRT models enabling parameter calibration in the presence of latent

classes are needed. Examining whether latent classes are present in international large-scale assessments based on the stated reasons is important for obtaining reliable and valid results.

Purpose

This study aims to determine the model which the data obtained from Singapore, Turkey, and South Africa countries that received booklet 7 of the 8th grade science subtest in TIMSS 2015 test, which is an international large-scale assessment, fits best in the presence of latent classes, thus contributing to the validity of the model. In this context, answers to the following research questions were sought;

(1) Which Mixture IRT (Rasch, 1PL, 2PL and 3PL) model do TIMSS 2015 science subtest items fit better for Singapore, Turkey and South Africa?
(2) What are the item parameters based on the model that fits best to data for Singapore, Turkey and South Africa?

Method

Study Group

The study group consists of 436 students from Singapore, 432 students from Turkey and 894 students from South Africa who attended TIMSS 2015 at the 8th grade level and were administered Booklet 7 science subtest. Table 1 shows the mean scores and standard deviations of the students for three countries.

Table 1. Descriptive statistics of the scores

	N	$\bar{X}$	SD
Singapore	436	12.41	3.90
Turkey	432	8.57	3.87
South Africa	894	5.12	2.52

The mean scores and standard deviations of the students from Singapore were calculated as 12.41 and 3.90, respectively. For the students from Turkey, the mean scores were 8.57 and the standard deviations were 3.87 while the mean scores of the students in South Africa were 5.12 and the standard deviations were 2.52.

Data Collection Tools

Different test booklets with common items are used to estimate student ability in international large-scale assessments (Xu, 2009). TIMSS 2015 had 14 different booklets organized according to the content domain and cognitive domain at the 4th and 8th grade levels. The 8th grade science subtest of TIMSS 2015 had four different content domains including physics, chemistry, biology and earth sciences as well as three cognitive domains of knowing, applying, and reasoning Each booklet was composed of similar proportions of item types, including multiple-choice and constructed-response items (Martin, et al., 2016). 18 multiple-choice items in booklet 7 of the science subtest were included in the analysis within the scope of this study. Correct answers were coded as 1 while wrong answers were coded as 0.

Data Analysis

The three-parameter Mixture IRT model including item parameters and the guess parameter for each class is shown with equation (1) (Choi, et al., 2015):

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

In this equation; $g=(1,2,..,G)$ indicates latent class membership for the three-parameter Mixture IRT model, (β_{ig}) indicates the interclass item difficulty parameter for item i, (α_{ig}) indicates interclass item discrimination for item i, (Υ_{ig}) indicates lower-asymptote, i.e. the chance parameter for item i, (θ_{jg}) indicates 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 (π_g) are estimated with the $\sum_{i=1}\pi_g=1$ and $0\leq\pi_g\leq1$ restriction (Rost, 1990; Sen et al., 2016). Mixture IRT models are nested models. That is because it turns into a Mixture 2-parameter model when the low-asymptote parameter is equal to zero, i.e. the chance is eliminated: the two-parameter Mixture IRT (Mix2PL) is shown with equation (2) (Finch & French, 2012):

$$P(x_{ij}=1|\theta_j)=P_{ij}=\sum_{g=1}^{G}\pi_g\left(\frac{exp[\alpha_{ig}(\theta_{jg}-\beta_{ig})]}{1+exp[\alpha_{ig}(\theta_{jg}-\beta_{ig})]}\right) \quad (2)$$

It is transformed into the 1-parameter model form with the assumption that the chance parameter is equal to zero and the item discrimination parameter is equal for all classes while it is transformed into the Mixture Rasch model form with the assumption that chance parameter is equal to zero and the item discrimination parameter is equal to 1. The formula of the Mixture Rasch model is shown by the following equation (3) (Rost, 1990):

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

In Mixture IRT models, the difficulty, discrimination, and guess parameters have the same meaning as the parameters in the overall IRT framework. Therefore, item difficulty provides information about the probability of an item to be answered correctly by the individual, discrimination indicates how well the item distinguishes between individuals with different levels of the measured construct, and the chance parameter is a measurement of the probability that the individual answers the item correctly by mere chance (de Ayala, 2009).

The Mixture IRT models were analyzed using the Mplus 7.4 program. In the Mplus program, parameter estimations are done using maximum likelihood (ML) and Bayesian estimation methods. Estimation is performed for missing data with the full information maximum likelihood (FIML) method by adding the "Missing All (99)" command (Muthen & Muthen, 2017). In this study ML method was used for parameter estimation and FIML method was used for missing data.

Model Fit

An exploratory approach that starts with a one-class solution and adds additional classes until obtaining the model that best fits the data is adopted for model fit in Mixture IRT models.

In one-class IRT models, both likelihood ratio tests and relative fit indexes can be used to determine the optimal model. On the other hand, the likelihood ratio test is not suitable for model comparisons between Mixture IRT models (Li, Cohen, Kim, & Cho, 2009; Nylund, Asparouhov, & Muthén, 2007). Using relative fit indices such as Akaike Information Criterion (AIC; Akaike, 1974), Bayesian Information Criterion (BIC; Schwarz, 1978), sample size adjusted BIC (SABIC; Sclove, 1987), and consistent AIC (Bozdoğan, 1987) is suggested for model-data fit in Mixture IRT models. However, simulation studies indicate that BIC tended to perform better between these indices (Nylund et al., 2007; Li et al., 2009; Sen, 2018). In this study, the BIC index was given for model-data fit and the AIC index was considered as the supporting index.

Label Switching

The parameters calibrated for Class 1 are sometimes labeled as Class 2 or vice versa as there is no information about the number and nature of the classes in Mixture models (McLachlan & Peel, 2000). This type of label switching can occur in Bayesian and ML estimations (Finch & French, 2012). As class labels are exchanged between data sets, parameter estimates to be collected over potentially mislabeled classes create an undesirable situation. In this case, the label switching problem can be solved by taking the estimated item parameter values as starting values. Model-data fit index values are not affected by Label Switching (Kutscher, Eid, & Crayen, 2019).

Findings

Table 2 shows the information criteria indices obtained from the analyses aimed at determining which Mixture IRT model (Rasch, 1PL, 2PL, and 3 PL) the data obtained from the science subtest for Singapore, Turkey, and South Africa that attended the 8th grade TIMSS 2015 fits best:

Table 2. Model data fit index values based on models

		Singapore		Turkey		South Africa	
		AIC	BIC	AIC	BIC	AIC	BIC
Mixture Rasch	Class 1	3938.306	3999.227	4497.233	4564.738	7667.356	7749.363
	Class 2	3858.520	3983.747	4208.711	4333.596	7038.496	7190.207
	Class 3	3953.595	4043.127	4205.932	4394.948	7010.274	7239.892
Mixture 1pl	Class 1	3931.087	3995.393	4499.149	4570.030	7641.918	7728.024
	Class 2	3860.354	3988.965	4200.266	4328.526	6978.372	7134.184
	Class 3	3859.465	4059.151	4204.579	4403.721	6969.846	7211.764
Mixture 2pl	Class 1	3917.386	4039.228	4433.041	4561.302	7579.963	7735.775
	Class 2	3860.552	4107.620	4201.933	4448.328	6950.774	7250.098
	Class 3	3862.277	4173.651	4204.076	4514.602	6921.683	7298.913
Mixture 3pl	Class 1	4049.960	4232.722	4505.475	4694.491	7096.437	7317.854
	Class 2	4051.924	4238.071	4341.081	4526.721	7274.629	7500.147
	Class 3	4053.923	4243.455	4343.081	4532.097	7276.630	7506.248

When AIC and BIC values are examined in general, it is observed that the AIC and BIC values for Singapore data are lower in the two-class Mixture Rasch model. Therefore, it can be said that the Singapore data fits the two-class Mixture Rasch model better. From a model-based point of view, it can be said that the two-class model fits the data better for the Mixture 1-parameter model and Mixture 2-parameter model while the one-class model fits the data better for the Mixture 3-parameter model based on the BIC indices.

As for the model-data fit indices for the Turkey data, it was also determined that AIC and BIC values were lower for the two-class Mixture 1-parameter model. Therefore, it can be said that the Turkey data fits the two-class Mixture 1-parameter model better. From a model-based point of view, it can be said that the two-class model fits the data better for the Mixture Rash model while the two-class model fits the data better for the Mixture 2-parameter model and Mixture 3-parameter model based on the BIC indices.

The model-data fit indices for the South Africa data indicate that the BIC value was lower for the two-class Mixture 1-parameter model while the AIC value was lower for the three-class Mixture 2-parameter model. Previous research has shown that the AIC index tends to select the model with a higher number of classes (Preinerstorfer & Formann, 2012; Sen, 2018). The lower AIC value for the three-class model is similar to other research results. In this context, it can be said that the South Africa data fits the two-class Mixture 1-parameter model better as the BIC index has a higher performance in terms of model-data fit (Li et al., 2009). From a model-based point of view, it can be said that the two-class model fits the data better for the Mixture Rasch model and the Mixture 2-parameter model while the one-class model fits the data better for the Mixture 3-parameter model based on BIC indices. As a result, it can be said that the data from Singapore fit the two-class Mixture Rasch model better while the data from Turkey and South Africa fit the two-class Mixture 1-parameter model better.

In an attempt to answer the second research question the item parameters obtained for the classes in the model that fits the data better in Singapore, Turkey, and South Africa, respectively, are provided in Table 3. Table 3 shows the item parameters estimated for the two-class Mixture models selected for the data obtained from these three countries.

Table 3. Item parameters calibrated for the two-class Mixture models

	Singapore			Turkey			South Africa		
	α	β1	β2	α	β1	β2	α	β1	β2
Item 1	1.00	-0.932	1.064	0.717	-0.203	2.960	0.446	2.847	5.987
Item 2	1.00	-1.667	-0.182	0.717	0.364	1.250	0.446	0.068	1.077
Item 3	1.00	-1.989	-0.730	0.717	0.213	0.468	0.446	-3.017	-0.202
Item 4	1.00	-0.329	0.105	0.717	1.699	2.593	0.446	1.294	4.148
Item 5	1.00	-1.817	-1.111	0.717	0.339	2.172	0.446	0.619	4.208
Item 6	1.00	-3.307	-1.347	0.717	-2.208	-0.247	0.446	-1.089	0.349
Item 7	1.00	-3.732	-2.663	0.717	-4.880	-1.804	0.446	-2.702	-0.363
Item 8	1.00	-2.849	-0.395	0.717	-3.820	-0.592	0.446	-0.670	1.299
Item 9	1.00	-1.282	-0.312	0.717	-1.062	0.925	0.446	-0.828	-0.359
Item 10	1.00	-9.481	0.954	0.717	-3.449	1.673	0.446	0.054	1.041
Item 11	1.00	0.338	0.967	0.717	0.505	2.961	0.446	1.720	3.373
Item 12	1.00	-4.245	-2.504	0.717	-4.354	3.571	0.446	0.608	3.245
Item 13	1.00	-2.616	-0.767	0.717	-3.414	0.593	0.446	0.241	3.391
Item 14	1.00	-0.850	0.427	0.717	-1.269	0.998	0.446	-0.633	2.530
Item 15	1.00	-1.249	-1.518	0.717	-2.123	-0.978	0.446	-0.728	-0.639
Item 16	1.00	-1.662	-0.463	0.717	-1.393	0.517	0.446	0.754	0.953
Item 17	1.00	-0.590	1.251	0.717	1.141	2.753	0.446	0.426	3.292
Item 18	1.00	-3.420	-2.025	0.717	-1.542	-2.119	0.446	-1.892	-0.589

The item difficulty (β1 and β2) and item discrimination parameters (α) obtained from two-class Mixture IRT models that better fit the data for Singapore, Turkey and South Africa are shown in Table 3. As the Singapore data fitted the Mixture Rasch model, the discrimination parameter

was estimated as 1 while the discrimination parameters for the data from Turkey and South Africa were estimated as 0.717 and 0.446 respectively since they fitted the Mixture 1-parameter model. The item difficulty averages for the first latent classes were estimated as 2.18, -1.48 and -0.16 respectively and for the second latent classes were estimated as -0.49, 0.89 and 1.87 respectively in Singapore, Turkey and South Africa data. When the estimated item difficulty parameters for the Singapore data are examined, it is observed that the item difficulty parameters in Class 1 vary between -9.481 (item10) and 0.338 (item 11), meaning that these items are usually very easy for students in Class 1. The item difficulty parameters of the items in Class 2 vary between -2.663 (item7) and 1.064 (item 1). As a result, it can be stated that the students in Class 2 had a slightly lower performance than the students in Class 1. The fact that the vast majority of the items in both classes have negative difficulty values could indicate that the items were very easy for the students in Singapore.

The analysis of the item difficulty parameters calibrated for the Turkey data reveals that the item difficulty parameters vary between -4.880 (item 7) and 1.699 (item 4) in Class 1 and between -2.119 (item 18) and 3.571(item 12) in Class 2. In this case, it can be said that the items were easier for students in Class 1 and the students in this class performed better while the items were a little harder for the students in Class 2 and the students in this class had a lower performance. When the item difficulty values in the latent classes in the Singapore and Turkey data are compared, it can be stated that the items were harder for the students in Turkey.

A label switching problem encountered in Mixture models was identified in the South Africa data. The analysis output revealed that the item parameters estimated for Class 1 were labeled as Class 2. This problem was solved by taking the estimated item parameter values as starting values (Kutscher et al., 2019). When item difficulty parameters are examined for South Africa data, it is observed that item difficulty parameters in Class 1 range from -3.017 (item 3) to 2.847 (item 1) while item difficulty parameters in Class 2 range from -0.639 (item 15) to 5.987 (item 1). In this case, it can be said that the items were usually a little easier for students in Class 1 and the students in this class performed slightly better while the items were a little harder for the students in Class 2 and the students in this class had a lower performance. When the item difficulty values in the latent classes in the Singapore and Turkey data are compared with the item difficulty values in the latent classes in the South Africa data, it is observed that the items were highly difficult for the students in South Africa. Percentages of students in latent classes for each country are given in Table 4:

Table 4. Percentages of examinees for countries by latent class (LC)

	Singapore	Turkey	South Africa
Latent Class 1	0.59	0.62	0.11
Latent Class 2	0.41	0.38	0.89

The conditional probability values for the latent classes given in Table 4 reveal that 59% of the students in Singapore were in Class 1, 41% were in Class 2, 62% of the students in Turkey were in Class 1, 38% were in Class 2, 11% of the students in South Africa were in Class 1 and 89% were in Class 2. The high percentage of underperforming students in the South African overlaps the fact that South Africa ranked last in the 8th grade science test of TIMSS 2015. According to student percentages in latent classes, the data obtained from the students who took the 8th grade science subtest of TIMSS 2015 has a heterogeneous structure consisting of two homogeneous subclasses. Therefore, this result shows that Mixture IRT models are needed to detect latent classes in TIMSS 2015 data.

Discussion and Conclusion

Standard IRT models are used for calibration of item parameters and scaling of individual performances in international large-scale assessments such as TIMSS and PISA (Martin et al., 2016). Literature review revealed that latent classes are ignored at the end of the analyses conducted with IRT models in some studies that employed international large-scale test data (Kreiner & Christensen, 2014; Oliveri & von Davier, 2011; Oliveri & von Davier, 2014; Park et al., 2016). In the presence of latent classes, the parameter invariance assumption of standard IRT models is violated and biased results can be obtained in item parameter calibrations (DeMars & Lau, 2011). In large-scale assessments, the invariance of item parameters is often tested within the context of DIF studies. In these studies, however, the existence of latent classes is not checked, and latent traits are often neglected (Park et al., 2015). Therefore, the analysis was carried out using Mixture IRT models which allow item parameters to vary among latent classes.

Mixture Item Response Theory models (Rasch, 1PL, 2PL and 3PL) analysis results showed that Singapore data fitted the two-class Mixture Rasch model better while Turkey and South Africa data fitted the two-class Mixture 1-parameter model better. Choi et al. (2015) found out that TIMSS 2007 mathematics data fitted the two-class 3-parameter Mixture IRT model best, Zhang et al. (2015) found out that the data obtained from Chinese in the mathematics subtest of PISA 2009 fitted the two-class Mixture Rasch model best, and Sen et al. (2016) found out that the data obtained from South Korea in the 8th grade mathematics subtest of TIMSS 2011 fitted the two-class Mixture Rasch model best. When these results are considered, it is seen that that the results of this study show similarity to those obtained by applying Mixture IRT models to large-scale test data such as TIMSS and PISA.

Two latent classes were identified in Singapore, Turkey and South Africa data. It was concluded that the students in the first latent class in Singapore, Turkey and South Africa data performed better in answering items than the students in the second latent class. It was concluded that the students in the second latent class in Singapore, Turkey and South Africa data had a lower performance in answering items than the students in the first latent class. These results indicate the presence of latent classes in the data of countries with high, medium and low performance regardless of country's performance ranking. The parameter invariance assumption, which is one of the assumptions of standard IRT models, is violated in the presence of latent classes (Park et al.,2016). As the parameter invariance assumption could not be met, it was concluded that the data obtained from the 8th grade science subtest of TIMSS 2015 fit Mixture IRT models. As a result, Mixture IRT models are needed for calibration on subgroups basis in TIMSS 2015 assessment. Accordingly, as it is stated in studies stressing the importance of meeting model assumptions (Goldstein, 2004; Grisay & Monseur, 2007; Kreiner & Christensen, 2014; Oliveri & von Davier, 2011), it will be possible to reach more accurate conclusions about determining the strengths and weaknesses of countries, reorganizing, improving, and evaluating education programs based on findings resulting from the collection of correct evidence about the validity of results obtained from large-scale assessments.

This study is conducted on dichotomous scored items in the 8th grade science subtest of TIMSS 2015. Researchers can perform Mixture IRT model analyses with polytomous scored items. Moreover, although it can be stated that students give a low or high performance in the classes obtained with Mixture IRT models, no account can be provided for the cognitive levels of TIMSS (knowing, applying and reasoning) in which students in these classes are successful. Researchers can obtain more detailed information about the formation of latent classes by

cognitive domain levels using the Confirmatory Mixture IRT model approach. Furthermore, DIF studies can be conducted with Mixture IRT models using large-scale test data such as TIMSS and PISA.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723. doi: 10.1109/TAC.1974.1100705

Bock, R.D., & Zimowski, M.F. (1997). *Multiple group IRT*. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*,133–148. doi:10.1111/j.1745-3984.2005.00007

Cook, L. (July, 2006). *Practical considerations in linking scores on adapted tests*. Keynote address at the 5th International Meeting of the International Test Commission, Brussels, Belgium.

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

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

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

DeMars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: how accurately detect who is responding differentially? *Educational and Psychological Measurement*, *71*, 597–616. doi: 10.1177/0013164411404221

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, *22*(4), 249-262. doi: 10.1111/j.1745-3984.1985.tb01062.x

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16*(1), 20–30. doi:10.1027//1015-5759.16.1.20

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. *Quality of Life Research*. New Jersey: Lawrence Erlbaum Associates, Inc.

Embretson, S. E. (2007). Mixed Rasch models for measurement in cognitive psychology. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: extensions and applications* (pp. 235-253). New York: Springer Verlag.

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. doi: 10.22237/jmasm/1335845580

Glück, J., & Spiel, C. (2007). Studying development via item response model: A wide range of potential uses. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 281-292). New York: Springer Verlag.

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 319–330. doi:10.1080/0969594042000304618

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33*(1), 69–86. doi: 10.1016/j.stueduc.2007.01.006

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

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT Models. *Educational Measurement*: *Issues and Practice, 8*(1), 35–41. doi:10.1111/j.1745 3992.1989.tb00313.x

Hong, S. (2007). Mixed Rasch modeling of the self-rating depression scale. *Educational and Psychological Measurement, 67*(2), 28299.doi:10.1177/0013164406292072

Kreiner, S., & Christensen, K. B. (2007). Validity and objectivity in health-related scales: Analysis by graphical loglinear Rasch models. In M. von Davier, & C. H. Carstensen (Eds.). *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 329-346). New York: Springer Verlag.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness: A new look at the PISA scaling model underlying ranking of countries according to Reading literacy. *Psychometrika*, *79* (2), 210–231. doi: 10.1007/s11336-013-9347-z

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. doi:10.3389/fpsyg.2019.02494

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. doi:10.1177/0146621608326422

Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology, 9*. doi: 10.3389/fpsyg.2018.01372

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*, 975-999. doi:10.1080/00273171.2010.533047

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. Zugriff am (Vol. 21).

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley.

Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure: A mixture-distribution Rasch analysis. *European Journal of Psychological Assessment, 24*(1), 27–34. doi:10.1027/1015-5759.24.1.27

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi:10.1007/BF02294825

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*,13–23. doi:10.3102/0013189X023002013

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195-215. doi:10.1007/BF02295283

Mislevy, R., & Huang, C., W. (2007). Measurement models as narrative structures. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 15-35). New York: Springer Verlag.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (Eighth Edition). Los Angeles, CA: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthe´n, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535-569. doi:10.1080/10705510701575396

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*. doi: 10.3389/fpsyg.2016

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.

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. doi:10.1080/15305058.2014.891223

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*(1), 1–21. doi:10.1080/15305058.2013.825265

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. doi:10.1111/j.2044-8317.2011.02020.x

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

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

Rost, J., & von Davier, M. (1993). *Measuring different traits in different populations with the same items*. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), Psychometric Methodology: Proceedings of the 7th European Meeting of the Psychometric Society in Trier (pp. 446-450).

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). New York, NY: Wax-Mann.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective*. Doctoral Dissertation. Available from ProQuest Dissertations and Theses database. (UMI No. 3175148)

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

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

Sen, S. (2018). Spurious latent class problem in the mixed Rasch model: A comparison of three maximum likelihood estimation methods under different ability distributions. *International Journal of Testing*,18(1), 71-100. doi: 10.1080/15305058.2017.1312408

Toker, T. (2016). *A comparision latent class analysis and the mixture Rasch model: A cross-cultural comparision of 8th grade mathematics achievement in the fourth international mathematics and science study (TIMSS-2011).* Doctoral Dissertation, The Faculty of the Morgridge College of Education University of Denver, USA.

von Davier, M., & Yamamoto, K. (2007). Mixture- distribution and hybrid Rasch models. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 99-115). New York: Springer Verlag.

Xu, Y. (2009). *Measuring change in jurisdiction achievement over time: Equating issues in current international assessment programs*. Doctoral Dissertation, University of Toronto.