International Journal of Assessment Tools in Education
2021, Vol. 8, No. 1, 67-89
https://doi.org/10.21449/ijate.705426
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Examining the Measurement Invariance of TIMSS 2015 Mathematics Liking Scale through Different Methods

Zafer Erturk [1,*], **Esra Oyar** [1]

[1]Gazi University, Department of Educational Science, Ankara, Turkey

ARTICLE HISTORY

Received: Mar. 17, 2020
Revised: Dec. 01, 2020
Accepted: Jan. 05, 2021

KEYWORDS

Measurement Invariance,
TIMSS,
Latent Class,
Mixed Rasch Model,
Factor Analysis

Abstract: Studies aiming to make cross-cultural comparisons first should establish measurement invariance in the groups to be compared because results obtained from such comparisons may be artificial in the event that measurement invariance cannot be established. The purpose of this study is to investigate the measurement invariance of the data obtained from the "Mathematics Liking Scale" in TIMSS 2015through Multiple Group CFA, Multiple Group LCA and Mixed Rasch Model, which are based on different theoretical foundations and to compare the obtained results. To this end, TIMSS 2015 data for students in the USA and Canada, who speak the same language and data for students in the USA and Turkey, who speak different languages, are used. The study is conducted through a descriptive study approach. The study revealed that all measurement invariance levels were established in Multiple Group CFA for the USA-Canada comparison. In Multiple Group LCA, on the other hand, measurement invariance was established up to partial homogeneity. However, it was not established in the Mixed Rasch Model. As for the USA-Turkey comparison, metric invariance was established in Multiple Group CFA whereas in Multiple Group LCA it stopped at the heterogeneity level. Measurement invariance for data failed to be established for the relevant sample in the Mixed Rasch Model. The foregoing findings suggest that methods with different theoretical foundations yield different measurement invariance results. In this regard, when deciding on the method to be used in measurement invariance studies, it is recommended to examine the necessary assumptions and consider the variable structure.

1. INTRODUCTION

In a world of rapid development and globalization, the information in the social, geographical, political, healthcare and educational fields of countries are easily accessed through a variety of organizations. An international database is thus possible because information regarding all countries is accessible. TIMSS -Trends in International Mathematics and Science Study and PISA - Programme for International Student Assessment are among the international educational databases. By way of these large-scale assessments, students from different educational systems can be compared for their both cognitive (e.g., mathematics, science

CONTACT: Zafer ERTÜRK ✉ zerturk35@gmail.com Gazi University, Department of Educational Science, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

achievement) and affective (attitude, perception, self-confidence, motivation, etc.) latent traits (Buchholz & Hartig, 2017).

There are a number of studies in the literature conducting cross-cultural comparisons thanks to the accessibility to international data (e.g. Alatlı, Ayan, Demir & Uzun, 2016; Asil & Gelbal, 2012; Rutkowski & Svetina, 2014). In international assessments such as TIMSS, data is collected by administering a single measurement instrument to all participants from different countries. However, people from different socio-cultural backgrounds are likely to have different social, ethical and value judgments and interpret the scale items differently from each other. Thus, when collecting data from individuals from different cultures, researchers need to ensure that the items in measurement instruments mean the same in every culture.

The measurements need to be valid to obtain accurate results from the group comparisons made using the same measurement instrument (scale, questionnaire, test, etc.). In TIMSS, individuals from different cultures are administered the same measurement instruments. Therefore, the original versions of these instruments are translated into the languages spoken in all countries. The fact that measurement instruments can be translated flawlessly into other languages does not guarantee that each culture interprets the questions in the same way (Kankaras, 2010). Thus, there is an increasing need for addressing the methodological problems arising from the comparison of the data obtained from different countries and different cultures. One of these problems in intercountry comparisons is the invariance of measurements. In this regard, one of the basic concerns in any cross-cultural studies is whether or not the measurement invariance is established in testing the differences among groups (Hui & Triandis, 1985). In their study, Arım and Ercikan (2014) examined to what extent TIMSS 1999 U.S. and Turkey mathematics test results are comparable. In the comparison of the two countries, measurement invariance was taken into account and changing item function analyzes were performed for this. Accordingly, in the analysis made by comparing the test characteristic curves, it was determined that approximately 23% of the mathematics items operate differently between these two countries.

1.1. Measurement Invariance

Bryne and Watkins (2003) defined measurement invariance as the perception and interpretation of the items in the measurement instrument in the same way by individuals who in different sub-groups with respect to a certain variable. Invariance of measurements and methods that are adopted in cross-cultural studies across groups is referred to as the methodological invariance. Scale invariance and item invariance indicate the methodological invariance and concentrate on the degree of similarity between measurement methods across cultures (Kankaras, 2010).

Measurement invariance is a proof of validity employed to show that the same measurement instrument which is administered to different cultures in a study measures the same construct. In addition, since measurement instruments are created to measure a specific construct, the participants' responses should reveal their position about that specific construct. If their responses are influenced by additional factors which are different across cultures aside from the aimed construct, the invariance of measurements will fail to be established. In this case, the results to be obtained about the individuals by means of the measurement instrument will not reflect the real scores.

In order to test whether or not measurement invariance, a prerequisite in international comparison studies, is established, Multiple Group Confirmatory Factor Analysis (MG-CFA), an extension of Structural Equation Modeling (SEM), and the methods under the Item Response Theory (IRT) are adopted (Eid, Langenheine & Diener, 2003). In addition to these methods, mixed distribution models in which measurement invariance is examined by way of identifying the heterogeneous sub-groups are also implemented. Mixed distribution models have been

developed for the Item Response Theory (Mislevy & Verhelst, 1990; Rost & von Davier, 1995; von Davier & Rost, 1995;) and Structural Equation Models (Yung, 1997). These methods are combinations of a latent trait or the latent class analysis and a structural equation model (Eid & Rauber, 2000). Multiple Group Latent Class Analysis (MG-LCA), which is a method among mixed distribution models and is dependent on latent class analyses, may also be employed in measurement invariance studies (Magidson & Vermunt, 2001; Moors, 2004; Moors & Wennekers, 2003). MG-CFA is the method which is used when the observed and latent variables are continuous but cannot be used when both are categorical (Somer, Korkmaz, Dural & Can, 2009). MG-LCA, on the other hand, which is covered by the latent class models, can be used in measurement invariance studies if the two data structures mentioned are categorical. In addition, in their study in which MG-CFA and Differential Item Functioning (DIF) are compared based on IRT and MG-LCA, Kankaras, Vermunt and Moors (2011) stated that MG-LCA was an excellent alternative to the other two methods.

Another mixed distribution model is the Mixed Rasch Model (MRM). In mixed distribution Rasch models (Rost & von Davier, 1995), latent classes may be formed under a Rasch model for all individuals in a population and item difficulty parameters may differ across the unknown sub-groups (latent classes). Using this methodology, the number of groups reguired to account for the differences in item parameters can be identified." In addition, the probability that an individual may belong to different classes can be calculated and individuals may be assigned to a latent class where their membership probability is maximum (Eid & Rauber, 2000). For ordered response categories (e.g., Likert-type scales), polytomous mixed Rasch model can be applied (Rost, 1991).

1.1.1. *Multiple Group Confirmatory Factor Analysis*

MG-CFA, a commonly preferred method in measurement invariance studies in various disciplines (Meredith, 1993; Mullen, 1995; Steenkamp & Baumgartner, 1998), is a parametric and linear approach investigating the similarity between measurement model parameters named as factor loadings, intercepts and error variances for the same factor models across groups.

Measurement invariance within the scope of MG-CFA is defined and tested through four hierarchical models (Byrne & Stewart, 2006; Meredith, 1993; Wu, Li & Zumbo, 2007). The measurement invariance levels that are tested in MG-CFA can be listed respectively as follows:

i. **Configural Invariance:** The configural model is the first level where the measurement invariance is tested in MG-CFA. This step allows freely estimating the factor loadings, regression constants and error variances concerning the groups.
ii. **Metric (Weak) Invariance:** Metric invariance, the second level, is the step where measurement units of groups regarding the latent variable are tested to find out whether they are similar or not. To this end, factor loadings are also restricted in addition to the factor number and factor pattern in groups.
iii. **Scalar (Strong) Invariance:** This model involves the restriction of the regression constants as well as the factor pattern and factor loadings (Tucker, Ozer, Lyubomirsk, & Boehm, 2006, p. 344).
iv. **Strict Invariance:** It is the last step of measurement invariance. The hypothesis that error terms concerning the items in the measurement invariance are equivalent across comparison groups is tested on this level (Önen, 2009).

There is a myriad of measurement invariance studies in Turkey conducted through MG-CFA. Based on TIMSS 1999 data for Turkey, Uzun and Öğretmen (2010) identified the affective factors that are influential in students' science achievement and tested these factors' measurement invariance by gender. In another study, Bahadır (2012) modeled the variables affecting students' reading skills by means of PISA 2009 data for Turkey. Then she tested the

measurement invariance of the obtained modelacross regions using MG-CFA. There are also studies which investigate the measurement invariance by gender and regions (Gülleroğlu, 2017; Ölçüoğlu 2015; Uzun, 2008) as well as those that compare the countries (Asil & Gelbal, 2012; Güzeller, 2011), by means of MG-CFA and based on the data on Turkey obtained from international assessments such as TIMSS and PISA. In his study, Güzeller (2011) examined whether the factor structure of the Computer Attitude Scale in PISA 2009 is similar across 10 different countries, in other words, its cross-cultural measurement invariance is made through MG-CFA. He obtained a similar factor structure as a result of the confirmatory factor analysis performed for all countries and showed that computer attitude has a cross-cultural invariance. Asil and Gelbal (2012) analyzed the cross-cultural and interlingual invariance of the student questionnaire administered within the scope of the Programme for International Student Assessment (PISA) 2016 comparatively based on the samples of Australia, New Zealand, USA and Turkey. In the conclusion part of their study, they stated that the measurement invariance failed to be established because of translation-related problems and cultural differences. Wu, Li, and Zumbo (2007) investigated the cross-country measurement invariance using TIMSS 1999 data in their study. Accordingly, by using the mathematics achievement scores of 21 countries participating in TIMSS 1999, it was checked whether the measurement invariance was achieved with MG-CFA. These countries include the U.S.A and Canada. According to the results obtained from the study, strict invariance was provided between the U.S.A and Canada.

In the study conducted by Bowden, Saklofske, and Weiss (2011), the invariance of the measurement models of the Weschler Adult Intelligence Scale in U.S.A and Canada samples were examined. The model met the subtest scores that reflect similar structure measurement in both country samples and the assumption of invariance between samples. The results showed that structural validity was ensured in the measurement of cognitive abilities in U.S.A and Canadian samples and emphasized the importance of local norms.

1.1.2. *Multiple Group Latent Class Analysis*

MG-LCA as a concept is similar to MG-CFA in that it examines the relationship between categorical variables and latent constructs. MG-LCA analyzes the categorical latent constructs under the categorically observed variables whereas MG-CFA and IRT assume that latent variables are continuous. MG-LCA models the latent constructs as ordered categorical or nominal. Thus, instead of using the correlation/covariance matrix of data as done by MG-CFA, MG-LCA analyzes the cross-classification of the responses concerning the relevant variable (Kankaras, 2010). Measurement invariance within the framework of the latent class model is defined as the situation where the individuals who belong to different groups but are in the same latent class have the same observed response pattern and conditional probabilities (Millsap & Kwok, 2004).

Whether observable behaviors of individuals, such as attitudes, self-confidence, interest, willingness to study, and expressing that they find the lesson fun, arise from a latent structure is examined with latent variable models. There are three basic variables in these models: latent, observed and error. Observed variables are predicted by error and latent variables which explain the relationship between the observed variables, but the observed variables are not the cause of the latent variable (Collins & Lanza, 2010). In other words, if there is a latent variable that can be defined, the relationship between the observed variables disappears and this relationship is explained by the latent variable or variables (Goodman, 2002). Various models are available according to the fact that the variables are continuous and discontinuous. In latent class analysis, latent and observed variables are discontinuous. Latent variables observed in a traditional Latent Class Analysis consist of data at categorical or nominal scale level.

The latent class has at least two classes, if a model that can be defined with a single class is obtained, the observed variables are statistically independent of each other, so no latent

variables can be defined. The size of latent classes gives researchers information about subgroups in the universe. Another parameter used in the latent class analysis is conditional probabilities. Conditional probabilities can be likened to factor loadings in factor analysis. These parameters indicate the probability that an individual / observation in the t class of the X latent variable is at a certain level of the observed variable. Like the latent class probabilities, the sum of the conditional probabilities equals 1 (McCutcheon, 1987).

The most prominent reason why MG-LCA is preferred in measurement invariance studies is that almost all of the questions covered by the studies contain discrete (categorical or ordinal) response categories and can be used to identify the latent constructs from within the set of discrete observed variables (Kankaras, Moors & Vermunt, 2010). In addition, unlike MG-CFA and multiple group IRT which have strong assumptions about the distribution of data, MG-LCA is a rather flexible method feasible for all types of data. Second, while MG-CFA necessitates the invariance of at least two items under each factor to establish at least partial validity, there is no such requirement in MG-LCA. MG-LCA allows comparisons between groups even though each response in the model cannot establish the measurement invariance of the variable (Kankaras, 2010). In MG-LCA, the measurement invariance is gradually compared based on three basic models:

i. Heterogeneous Model: In this model, which is tested in the latent class analysis on the first level of measurement invariance, parameters to be estimated (conditional or latent class probabilities) are not restricted. In other words, each parameter is allowed to be estimated separately in comparison groups (McCutcheon & Hagenaars, 1997).
ii. Partial Homogeneous Model: Partial homogeneous model is the model in which slope parameters are tested by restriction. In this model, whether or not latent class probabilities differ across groups can be examined by way of removing the group-latent variable interaction effect from the model (Kankaras, Moors & Vermunt, 2010).
iii. Homogeneous Model: This is the next step after the partial homogeneous model is tested. The homogeneous model step in the Latent Class Analysis is equivalent to the scalar (strong) invariance model in the structural equation modelings and fixed parameters are also restricted in addition to the slope parameters.

There are also measurement invariance studies carried out through MG-LCA in Turkey (Güngör, Korkmaz & Somer, 2013; Yandı, Köse & Uysal, 2017). Güngör, Korkmaz and Somer (2013) carried out a study which examined the measurement invariance by gender of the love capacity dimension of Values in Action Inventory through MG-LCA. They obtained two latent classes for both men and women and established the homogeneous model among the measurement invariance steps. In their study, Yandı, Köse and Uysal (2017) compared measurement invariance results acquired from the models having different statistical assumptions. In the data obtained from the Openness for Problem Solving Scale in PISA 2012, when the measurement invariance is examined through the invariance of mean covariance structures analysis having the assumption of normality, the steps up to strict invariance were accepted whereas, in MG-LCA, which does not require the assumption of normality, the partial homogeneous model was accepted.

1.1.3. *Mixed Rasch Model*

MRM is the combination of the Rasch model and the latent class analysis (Rost, 1991). In MRM, the probability of answering correctly is a function of both the individual's skill, which is a continuous variable and the individual's group, which is a categorical variable. The standard unidimensional Rasch model assumes that the responses or answers to the items of individuals who are at the same skill level have the same response technique (Fischer & Molenaar, 2012). Thus, the estimation of item difficulty to be obtained from the analyses remains constant across

different latent groups at the same skill level (Baghaei & Carstensen, 2013). If the measurement invariance in a dataset having two or more latent classes is examined through the standard Rasch model, the results may be misleading for the researcher since they will be interpreted based on a single class (Frick, Strobl & Zeileis, 2015).

In the mixed Rasch model, first, the number of the latent classes is identified in the examination of the measurement invariance. The formation of a single latent class is interpreted as the establishment of measurement invariance. If more than one class is formed, the establishment of measurement invariance is said to fail and effort is made to find out whether an item-based Differential Item Functioning (DIF) is present or not. (Yüksel, 2015). DIF is the case where individuals from different groups but at the same θ level are not likely to give the same answer to an item. A DIF investigation involves the comparison of the differences between item difficulties in different latent classes. Researchers argued that interpreting the response patterns of the individuals in each latent class would be more efficient than attempting to define the latent classes formed through MRM by the observed groups at hand (Bilir, 2009; Cho, 2007; Cohen & Bolt, 2005). In addition, Kelderman and Macready (1990) stated that approaching the DIF problem through MRM is more advantageous. The Mixed Rasch Model can be used in the analysis of the tests measuring the affective traits as well as in the achievement tests (Rost, Carstensen & von Davier, 1997).

Many studies tested the measurement invariance by means of MRM (Aryadoust, 2015; Aryadoust & Zhang, 2016; Cohen & Bolt, 2005; Eid & Rauber, 2000; Pishghadam, Baghaei & Seyednozadi, 2017; Şen, 2016; Yalçın, 2019; Yüksel, 2015; Yüksel, Elhan, Gökmen, Küçükdeveci & Kutlay, 2018). Tee and Subramaniam (2018) analyzed the measurement invariance of the attitudes towards eighth grade science in the UK, Singapore and USA countries that entered TIMSS 2011 with Rasch analysis. According to the results obtained from the research, there are some differences between students in Asia and students in the West. More specifically, Singaporean students acknowledge the instrumental value of science more than students in the UK and the US. Although Singaporean students are more successful than students from the USA and the UK, they are less confident in science. When it comes to their feelings for sicence, again, Singaporean students love science more than U.S.A and U.K students.

Ölmez and Cohen (2018) in their study, Partial Credit Model of Mixed Rasch Models of the sixth and seventh grade students in Turkey are used to identify differences in mathematics anxiety. Two latent classes were identified in the analysis. While students in the first latent class have less anxiety about understanding mathematics lessons and the use of mathematics in daily life, students in the second class have more self-efficacy for mathematics. Students in both classes are similar in terms of exam and assessment anxiety. In addition, it was observed that students in the first latent class were more successful in mathematics, mostly liked mathematics and mathematics teachers, and had better-educated mothers than students in the second latent class. In addition, observed variables such as gender, private or public school attendance, and education levels of fathers did not differ significantly between latent classes.

1.2. Purpose

Measurement instruments are created based on the assumption that "an instrument measures the same construct in each group" (Başusta & Gelbal, 2015). The results of the studies in which the measurement invariance of the measurement instruments administered to different groups and different cultures remains untested may raise a lot of question marks in minds. Thus, the invariance of the measurement instruments needs to be tested before the initiation of intergroup, intercountry or cross-cultural comparisons. Since testing the measurement invariance makes a significant contribution to the validity of the results in comparison studies, the selection of the method to be utilized in compliance with the data structure when testing the measurement

invariance and fulfillment of the assumptions are of such importance. Thus, the validity of measurements would be further proved as the researchers adopt various methods to test the measurement invariance (Kankaras, Vermunt & Moors, 2011).

The purpose of this study is to investigate the measurement invariance of the data obtained from the "Mathematics Liking Scale" in TIMSS 2015 through MG-CFA, MG-LCA and MRM, which are based on different theoretical foundations and compare the obtained results. To this end, the country level was taken into consideration when forming the sub-groups. Mathematics achievement rankings were taken into account when determining the 3 countries included in the study. Comparisons were made between America, which is in the middle in the success ranking, and Canada, which is more successful. The analysis was also made between Amerika and Turkey which is less successful. In addition, the measurement invariance between the countries where the same language is spoken (USA and Canada) and the countries where different languages are spoken (USA and Turkey) was tested.

In this study, the Mixed Rasch Model, which is one of the methodologically prominent Mixed Item Response Theory models in test development and measurement invariance studies, and MG-LCA model and MG-CFA methods are focused on. The comparison of KRM and MG-LCA methods, whose mathematical methodologies are similar, will provide guiding results for researchers who will use these methods. In addition to the KRM and MG-LCA methods, the MG-CFA method, which has been used in measurement invariance studies for many years, was included in the study, and the validity of the study results was increased. In this study, the theoretical foundations of analysis methods used in the field of measurement invariance are explained in detail. In addition, testing the linguistic measurement invariance will also provide us with more valid information about the significance of the comparisons made according to cultural differences in the TIMSS 2015 student survey.

2. METHOD

2.1. Research Design

The purpose of this study is to investigate the invariance of the "Mathematics Liking Scale" in TIMSS 2015 in American, Canadian and Turkish cultures through MG-CFA, MG-LCA and MRM. The current research is a descriptive study as aims to identify the cross-cultural validity level of the "Mathematics Liking Scale" in TIMSS 2015study (Karasar, 2013).

2.2. Population and Sample

6079, 8068 and 9509 eighth-grade students from Turkey, Canada and the USA, respectively, participated in the TIMSS 2015 developed by the International Association for the Evaluation of Educational Achievement (IEA). A two-step path is pursued in the sample selection for TIMSS 2015. In this process, the schools are first selected from both public and private schools in each country through random sampling. Afterward, a class is chosen from each school (Olson, Martin & Mullis, 2008). The reason why eighth grade students were chosen in the study is that students' interests and attitudes towards mathematics are more pronounced in this age range. Since eighth grade students are in the last grade of primary education, they know themselves better than fourth grade students and their interests and attitudes towards lessons do not change much.

2.3. Data Collection Tool

The Mathematics Liking Scale in TIMSS 2015, which aims to identify whether or not students like math class, consists of a total of 9 items (TIMSS, 2015). Items were translated into Turkish by the researchers. The reason for using the " Mathematics Liking" scale within the scope of the study is the high number of items. In addition, the "Mathematics Liking" scale reflects general affective expressions towards mathematics. Thus, the perception of the statements in

the items is similar for the students of each country participating in TIMSS. The items are presented in the Appendix Table A1 both in English and in Turkish with their codes.

2.4. Data Analysis Procedures

The study employed three different methods, namely Multiple Group CFA, Multiple Group LCA and Mixed Rasch Model, in testing the measurement invariance. The steps followed in the analysis of data are as follows:

i. Calculation of the required statistics for the missing data, extreme value, normality, homogeneity of variance and multi-variant normality (testing of assumptions).
ii. Performance of CFA
iii. Performance of MG-CFA and testing of the levels of measurement invariance
iv. Performance of Latent Class Analysis and testing of the levels of measurement invariance
v. Implementation of the MRM and examination of the results
vi. Comparison of the methods based on the obtained results

2.4.1. *Assessment Criteria*

The MG-CFA method involves calculating the differences between the CFI and TLI values in comparing the two models in order to find out whether the measurement invariance is established. Measurement invariance is not established when ΔCFI and ΔTLI values are below -0.01 or above 0.01 (Byrne, Shavelson & Muthen, 1989; Li, Wang, Shou, Zhoung, Ren, Zhang & Yang, 2018; Liang & Lee, 2019; Schnabel, Kelava, Van de Vijver & Seifert, 2015; Wu, Li & Zumbo, 2007).

In the LCA model selection process, the simplest (parsimony) model, in other words, the model having the least number of latent classes and in which less parameter is predicted is sought. Statistical criteria, parsimony and interpretability should be considered in the model selection process (Collins & Lanza, 2010; Silvia, Kaufman & Pretz, 2009). There are several criteria in MG-LCA that are frequently used in the assessment of model-data fit. The likelihood ratio chi-square (L2) statistics are used as a standard criterion for the inconsistency between the observed and expected frequencies in the model. In addition to L2 statistics, various information criteria including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), modified AIC (AIC3) and consistent AIC (CAIC) are used in testing the measurement invariance in MG-LCA. When the sample size is large, BIC and CAIC are used for the model-data fit. When the sample size is small or medium, however, usually AIC statistics is used (Kankaras, Moors & Vermunt, 2011).

In order to identify the appropriate model-data fit in Mixed Rasch Model, aside from the criteria such as AIC and BIC as in MG-LCA, different statistics may be used, for example, the significance levels of Cressie Read and Pearson Chi-square values. Accordingly, the model obtained when p-value of Cressie Read or Pearson Chi-square is equal to or above 0.05 is said to be the appropriate model (von Davier, 2000). In addition, a common problem concerning chi-square parameters for the scale data observed in item-response models is that the number of cells significantly greater than the number of response models. The bootstrap method is recommended as a solution to this problem (Langeheine, Pannekoek & van der Pol, 1996). Thus, bootstrapped p-values of Cressie Read and Pearson Chi-square values are employed in this study to decide the appropriate number of latent classes.

In the event that a 1-class model is selected as the most appropriate model in model-data fit in MRM, it can be said that measurement invariance has been established, in other words, Differential Item Functioning (DIF) is not present in any of the items. However, if model-data fit cannot be ensured for a 1-class model, some items will be understood to have DIF. In testing DIF in items, item difficulties are calculated for the items in each class starting from the 1-class model to the latent class where the most appropriate model is identified. Identification of the

items displaying DIF involves the comparison of the differences between the item difficulty indices calculated for each latent class (Yüksel, Elhan, Gökmen, Küçükdeveci & Kutlay, 2018). Finally, contingency table analysis is performed to investigate whetherthe latent classes and observed variables (age, gender, status, country, etc.) are interrelated to find out the source of DIF occurring in some items.

2.4.2. *Testing of the Assumptions*

Items were reverse-coded as required before the pre-analysis. The missing data were removed from the dataset and excluded from the analyses. Deletion is preferred for the missing data, as it is not more than 5% in data and has a sufficient sample size. The testing of the assumptions was continued with 9509, 8068 and 5741 student data from the USA, Canada and Turkey, respectively.

In examining the extreme values, z score concerning the total scale score was calculated separately for each country and the values obtained were observed to be in the range of -1.54 and +1.95. In this regard, the data contained no extreme value. Skewness and kurtosis values were examined in testing the normality. Values for skewness and kurtosis were found to be in the range of ±1 for the entire group and for each country. Thus, the data were proved to fulfill the coefficient of normality (Büyüköztürk, 2017). In the analysis, LISREL 8.80 for MG-CFA; LATENT GOLD 5.0 for MG-LCA and WINMIRA 2001 package programs for MRM were used.

2.4.3. *Confirmatory Factor Analysis Results*

Firstly, in order to identify whether or not the measurement model developed in each step of the measurement invariance test established model-data fit, was performed and the obtained fit indices were reported and interpreted. CFA results for each country are shown in Table 1.

Table 1. *Model Fit Indices of Each Country Obtained from Measurement Models*

Fit Index	Measurement Model Results			Perfect Fit	Acceptable Fit
	US	Canada	Turkey		
RMSEA	0.09	0.08	0.10	0.00≤ RMSEA ≤ 0.05	0.05 ≤ RMSEA ≤ 0.10
CFI	0.98	0.99	0.98	0.97 ≤ CFI ≤ 1.00	0.95 ≤ CFI ≤ 0.97
TLI	0.98	0.98	0.97	0.95 ≤ TLI ≤ 1.00	0.90 ≤ TLI ≤ 0.95
NFI	0.98	0.99	0.98	0.95 ≤ NFI ≤ 1.00	0.90 ≤ NFI ≤ 0.95
AGFI	0.88	0.91	0.88	0.90 ≤ AGFI ≤ 1.00	0.85 ≤ AGFI ≤ 0.90
GFI	0.93	0.95	0.93	0.95 ≤ GFI ≤ 1.00	0.90 ≤ GFI ≤ 0.95

Table 1 shows that, based on the results of the measurement models developed separately for each country, the RMSEA values are in the acceptable range (Hooper, Coughlan & Mullen, 2008; Kelloway, 1989; Steiger, 1990)) while CFI, TLI and NFI values are in the perfect fit range (Sümer, 2000). AGFI and GFI values display perfect fit in the measurement model developed for Canada (Anderson & Gerbing, 1984; Cole, 1987) and are in the acceptable range for the USA and Turkey.

3. RESULT / FINDINGS

In this section, findings concerning MG-CFA, MG-LCA and MRM, which were employed to test the measurement invariances of the models obtained from the countries matched with respect to language (the same language or different languages) are presented.

3.1. Findings Obtained from MG-CFA

The results of MG-CFA that was performed to test the measurement invariance of data for "Mathematics Liking Scale" are presented in Table 2.

Table 2. *MG-CFA Results for USA-Canada and USA-Turkey Data*

	Steps	χ^2	sd	CFI	GFI	TLI	RMSEA	ΔCFI	ΔTLI
US-Can.	Configural Invariance[1]	4003.86	51	0.99	0.98	0.98	0.094	--	--
	Metric (Weak) Invariance[2]	4136.61	60	0.98	0.98	0.98	0.088	-0.01	0.00
	Scalar (Strong) Invariance[3]	4647.68	68	0.98	0.99	0.98	0.088	-0.01	0.00
	Strict Invariance[4]	5070.21	77	0.98	0.98	0.98	0.086	-0.01	0.00
USA-Tur.	Configural Invariance[1]	3714.90	50	0.98	0.98	0.97	0.098	--	--
	Metric (Weak) Invariance[2]	4064.85	60	0.98	0.97	0.98	0.094	0.00	0.01
	Scalar (Strong) Invariance[3]	6429.77	69	0.97	0.97	0.97	0.110	-0.01	0.00
	Strict Invariance[4]	7918.75	78	0.96	0.94	0.96	0.115	-0.02	-0.01

[1] Factor loadings, factor correlations and error variances are free
[2] Factor loadings are fixed (factor correlations and error variances are free)
[3] Factor loadings and factor correlations are fixed (error variances are free)
[4] Factor loadings, factor correlations and error variances are fixed

It is seen in Table 2 that model-data fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the measurement model developed in the configural invariance step given under USA-Canada comparison show a perfect fit. Therefore, it can be argued that the measurement model is the same for both countries. Fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the model developed in the metric invariance step display that the model-data fit is perfect. Examination of the difference between CFI and TLI values suggests that the difference is in the range of ±0.01 (ΔCFI = -0.01, ΔTLI = 0.00) and metric invariance is established. Fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the measurement model created to test the scalar invariance show that model-data fit is established. Examination of ΔCFI and ΔTLI reveals that the values are in the range of ±1 (ΔCFI = -0.01, ΔTLI = 0.00) and scalar invariance is established. Finally, model-data fit is seen to be established when the fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) in the developed strict variance model are examined. Examination of ΔCFI and ΔTLI reveals that the values are in the range of ±1 (ΔCFI = -0.01, ΔTLI = 0.00) and strict invariance is established. In conclusion, as a result of the analyses performed based on data on the USA and Canada, all steps of measurement invariance have been observed to be established.

Comparison of USA-Turkey samples shows that the model-data fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the measurement model which was developed to test the configural invariance reflect a perfect fit. Fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the model which was developed in the metric invariance step suggest perfect model-data fit. The difference between ΔCFI and ΔTLI values is shown to be in the range of ±0.01 (ΔCFI = 0.00, ΔTLI = 0.01). Fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) of the measurement model created to test the scalar invariance show that model-data fit is established. The ΔCFI and ΔTLI values are observed to be in the range of ±1 and the scalar invariance is established (ΔCFI = -0.01, ΔTLI = 0.00). Finally, model-data fit is seen to be established when the fit indices (RMSEA < 0.10, CFI > 0.95, GFI > 0.95, TLI > 0.95) in the developed strict variance model are examined. Examination of ΔCFI and ΔTLI values reveals that ΔTLI value is in the range of ±0.01 whereas ΔCFI is out of this range (ΔCFI = -0.02, ΔTLI = -0.01). In this case, strict invariance cannot be established. In brief, the results of the analyses performed based on the data on the USA and Turkey indicate that among the measurement invariance steps, configural, metric and scalar invariances are established but strict invariance cannot be established.

3.2. Findings Obtained from MG-LCA

In order to test the measurement invariance through MG-LCA, first, the number of latent classes is identified for Turkey, USA and Canada. The obtained statistics starting from 1 up to the 4-class model are examined to identify the number of latent classes in countries. The number of latent classes obtained for each country and the assessment criteria for classes are provided in Table 3.

Table 3. *Latent Classes and Information Criteria Values by Countries*

	Estimated number of parameters	*sd*	L^2	BIC	AIC	AIC3	CAIC
Turkey							
1-class	9	5732	78573.875	28961.187	67109.875	61377.875	23229.187
2-class	19	5722	73184.813	23658.679	61740.813	56018.813	17936.679
3-class	**29**	**5712**	72997.478	**23557.898**	**61573.478**	**55861.478**	**17845.898**
4-class	39	5702	**72997.477**	23644.451	61593.477	55891.477	17942.451
5-class	49	5692	72997.478	23731.005	61613.478	55921.478	18039.005
Canada							
1-class	9	8059	108154.927	35658.895	92036.927	83977.927	27599.895
2-class	19	8049	102234.857	29828.782	86136.857	78087.857	21779.782
3-class	**29**	**8039**	**102043.556**	**29727.438**	**85965.556**	**77926.556**	**21688.438**
4-class	39	8029	102043.556	29817.395	85985.556	77956.556	21788.395
5-class	49	8019	102043.556	29907.351	86005.556	77986.556	21888.351
US							
1-class	9	9500	135863.083	48843.140	116863.083	107363.083	39343.140
2-class	19	9490	126847.106	39918.763	107867.106	98377.106	30428.763
3-class	**29**	**9480**	**126565.792**	**39729.049**	**107605.792**	**98125.792**	**30249.049**
4-class	39	9470	126565.762	39820.649	107625.762	98155.792	30350.649
5-class	49	9460	126565.793	39912.249	107645.793	98185.793	30452.249

Table 3 shows that the three-class model has the lowest values for L2, BIC, AIC, AIC3 and CAIC in each country. In this context, it can be said that the latent variable of liking mathematics has three latent classes for the research sample. During the testing of the measurement invariance, analyses were performed based on the three-class model. Accordingly, first, the heterogeneous model, in which fixed and slope parameters are freely estimated, then, the partial homogeneous model in which slope parameters in both datasets are accepted equal and finally, the homogeneous model in which fixed parameters are also equalized in addition to slope parameters were created. First, the measurement invariance between the USA and Canada, where the same language is spoken, was tested. Accordingly, MG-LCA results for the USA-Canada sample are as shown in Table 4.

Table 4. *MG-LCA Results Obtained for the USA – Canada and USA- Turkey*

	Steps	Estimated number of parameters	*sd*	L^2	BIC	AIC	AIC3	CAIC
USA-Canada	Heterogeneous Model	166	17411	**57088.965**	-113092.181	**22266.965**	**4855.965**	-130503.181
	Partial Homogeneous Model	112	17465	57446.451	**-113262.510**	22516.451	5051.451	**-130727.510**
	Homogeneous Model	85	17492	58554.107	-112418.761	23570.107	6078.107	-129910.761
USA - Turkey	Heterogeneous Model	166	15084	**52902.889**	**-92391.248**	**22734.889**	**7650.889**	**-107475.248**
	Partial Homogeneous Model	112	15138	53877.805	-91936.478	23601.805	8463.805	-107074.478
	Homogeneous Model	85	15165	58080.332	-87994.024	27750.332	12585.332	-103159.024

Based on the comparison of the USA and Canada samples, it can be said that the most appropriate model according to BIC and CAIC is the partial homogeneous model (Kankaras & Moors, 2011). Comparison of USA-Turkey reveals that BIC and CAIC values are the lowest for the heterogeneous model. Thus, concerning the MG-LCA results for the USA-Turkey sample it can be said that the measurement invariance cannot be established.

3.3. Findings Obtained from MRM

In order to test the measurement invariance through MRM, first, the most appropriate number of latent classes to establish model-data fit for the USA-Canada and USA-Turkey were set. 400 bootstrap samples were used in each analysis to decide the number of the appropriate latent classes. The appropriate number of classes is decided considering the biggest insignificant p-value of Bootstrap Pearson χ^2 above 0.05. The number of latent classes and fit assessment criteria for the samples of USA-Canada and USA-Turkey are shown in Table 5.

Table 5. *Fit Statistics for the Mixed Rasch Model*

	Estimated number of parameters	BIC	Geometric Mean LL	Cressie Read (Bootstrap p-value)	Pearson χ^2 (Bootstrap p-value)
USA-Canada					
1-class	28	313816.58	0.37120018	0.000	0.000
2-class	57	301013.13	0.38687637	0.000	**0.097**
3-class	86	297434.88	0.39162740	0.000	0.010
USA-Turkey					
1-class	28	283209.56	0.35674063	0.000	0.000
2-class	57	269961.05	0.37476185	0.000	0.008
3-class	86	266247.67	0.38025276	0.000	0.022
4-class	115	264500.81	0.38306995	0.003	**0.500**
5-class	144	263102.19	0.38541874	0.000	0.013

According to the model assessment criteria in Table 5, one-class models in both samples, USA-Canada and USA-Turkey, are not appropriate. In this case, it can be claimed that the measurement invariance is not established for both samples. Once the establishment of the measurement invariance is failed, the appropriate number of classes to establish the model-data

fit is tried to be set. In the USA-Canada sample, in which the same language is spoken, only the p-value for Bootstrap Pearson χ^2 value of the two-class model is not significant ($p > 0.05$). In this case, the 2-class model was decided to be the most appropriate model for the USA-Canada sample. In the USA-Turkey sample, in which different languages are spoken, it is the four-class model in which the Bootstrap Pearson χ^2 value is not significant ($p > 0.05$).

Since the measurement invariance could not be established, item-based measurement invariance in MRM was examined. In this regard, first, the measurement invariance of nine items in the Mathematics Liking Scale was examined in the USA-Canada sample. The model establishing the model-data fit for the USA-Canada sample is the two-class model.

As for DIF, it emerges when differences take place between the difficulty parameters in classes. Item difficulty parameters obtained for each class are shown in Table 6. Comparison of the classes between rows allows identifying the items which are disproportionately easy or difficult and thus coming up with a clearer interpretation of each class.

Table 6. *Item Difficulty Estimations for Two-Class Model in the USA-Canada Sample*

Items	Class 1	Class 2
Item 1	0.949	0.408
Item 2	0.062	0.048
Item 3	-0.482	-0.138
Item 4	1.147	0.602
Item 5	0.546	0.233
Item 6	-0.600	-0.514
Item 7	-0.213	-0.155
Item 8	-0.769	-0.382
Item 9	-0.639	-0.102

Based on Table 6, Item 1 and Item 4 in the Latent Class 1 can be said to be more difficult than those in the Latent Class 2, in other words, individuals who are in Class 2 like mathematics less compared to the individuals in the Latent Class 1. On the other hand, it is seen that Item 8 and Item 9 are more difficult for the Latent Class 2, in other words, individuals who are in Class 1 like mathematics less compared to the individuals in the Latent Class 2. Some items were identified to have DIF as a result of the differentiation of difficulty parameters related to them into two latent classes. χ^2 test statistics is adopted to find out the source of DIF. Accordingly, since this study employs students from different countries, χ^2 analysis is performed between the students' latent classes and countries. 54% and 46% of the USA-Canada sample are made up of American and Canadian students, respectively. Results of the χ^2 test analysis performed between countries and class membership are shown in Table 7.

Table 7. *Results of χ^2 Analysis Between Latent Classes and Countries*

	Latent Class				
Country	1	2	Total	χ^2	p
U.S.A	5154 (54.2%)	4355 (45.8%)	9509 (54%)		
Canada	4985 (61.8%)	3083 (38.2%)	8068 (46%)	102.90	0.00*
Total	10139 (57.7%)	7438 (42.3%)	17577 (100.0%)		

* $p \leq .05$

Table 7 suggests a significant relationship between students' coming from different countries and latent class membership (χ^2 =102.90; $p \leq 0.05$). In this regard, DIF is considered to arise from students' coming from different countries. The rates of the American and Canadian students in Latent Class 1 are 54.2% and 61.8%, respectively. The rates in the second latent class are 61.8% for American students and 38.2% for Canadian students.

The measurement invariance of nine items in the Mathematics Liking Scale was examined for the USA and Turkey, where different languages are spoken. The model establishing the model-data fit for the USA-Turkey sample is the four-class model. Since a four-class construct emerged in the USA-Turkey sample speaking different languages, the measurement invariance could not be established. In this regard, in order to identify which items in the Mathematics Liking Scale prevent the measurement invariance from being established, in other words, display DIF, item difficulty parameters for each class were calculated and are presented in Table 8.

Table 8. *Item Difficulty Estimations for Four-Class Model in the USA-Turkey Samples*

Items	Class 1	Class 2	Class 3	Class 4
Item 1	1.066	0.711	0.872	0.264
Item 2	0.559	-1.041	-0.717	0.921
Item 3	-0.004	-0.742	-0.705	0.775
Item 4	1.218	1.525	0.584	-0.399
Item 5	0.696	0.359	1.002	0.093
Item 6	-2.021	-0.240	0.214	-0.461
Item 7	-0.577	0.058	0.181	-0.218
Item 8	-0.596	-0.472	-0.818	-0.637
Item 9	-0.340	-0.157	-0.613	-0.338

Examination of Table 8 reveals that item difficulty parameter values of the Latent Class 4 for Item 1, Item 4 and Item 5 are lower than the item difficulty values in other latent classes. Difficulty indices of the Latent Class 2 and the Latent Class 3 for Item 2 are observed to reflect quite low values as opposed to the difficulty indices of the Latent Class 1 and the Latent Class 4, which display very high values. For Item 3, the value of the difficulty parameter of the Latent Class 4 is much higher than that of the other latent classes. For Item 6, the item difficulty parameter value of the Latent Class 1 is much lower when compared to the other latent classes.

Considering that the difficulty parameters for some items are very different across four latent classes, the items can be claimed to have DIF. In MRM, χ^2 test statistics are used to identify the DIF source. The χ^2 analysis is carried out between the students' latent classes and countries in order to examine whether or not there is DIF with respect to coming from countries speaking different languages. 62% and 38% of the USA-Turkey sample are made up of American and Turkish students, respectively. Results of the χ^2 analysis performed between countries and class membership are shown in Table 9.

Table 9. *Results of χ^2 Analysis Between Latent Classes and Countries*

	Latent Class						
Country	1	2	3	4	Total	χ^2	p
U.S.A	4324	3146	1311	728	9509		
	(45.5%)	(33.1%)	(13.8%)	(7.7%)	(62%)		
Turkey	992	2560	1641	548	5741	1,363.13	0.00*
	(17.3%)	(44.6%)	(28.6%)	(9.5%)	(38%)		
Total	5316	5706	2952	1276	15250		
	(34.9%)	(37.4%)	(19.4%)	(8.4%)	(100.0%)		

* $p \leq 0.05$

Table 9 suggests a significant relationship between students' coming from different countries and latent class membership (χ^2=1363.13; $p \leq 0.05$). In this regard, students' coming from different countries can be suggested as a DIF source. The rates of American and Turkish students in Latent Class 1 are 45.5% and 17.3%, respectively. The rates in the second latent class are 33.1% for American students and 44.6% for Turkish students. The rates of American and Turkish students in Latent Class 3 are 13.8% and 28.6%, respectively. The rates in Latent Class 4 are 7.7% for American students and 9.5% for Turkish students.

4. DISCUSSION and CONCLUSION

Cross-cultural studies enable us to explore the universality of social and psychological laws and the cultural differences in people's characteristics, views and behaviors. A number of generalizations are made through comparison studies regarding the differences between the cultural groups. Thus, the validity of the results of the cross-cultural comparisons gains importance. Proving the validity comparison results necessitates testing the measurement invariance of measurement instruments because although the original measurement instrument can be translated into the languages of other cultures "flawlessly", it is not possible for each culture to interpret the questions in the same way (Hui & Triandis, 1985).

This study aims to examine the measurement invariance of the data obtained from the "Mathematics Liking Scale", which was administered to the students in TIMSS 2015 assessment by means of different methods, in countries, speaking the same and different languages. To this end, MG-CFA, MG-LCA and MRM methods which have different theoretical foundations were adopted.

As a result of the study, all steps of the measurement invariance was established when MG-CFA was employed for the analyses performed for the USA and Canada where the same language is spoken. In other words, data on these countries are comparable. When the measurement invariance was examined using the same data, it was seen that partial homogeneity was achieved by the MG-LCA. This step corresponds to the metric invariance in MG-CFA. In the MRM, another method used in the study, the measurement invariance for the USA-Canada sample could not be established and some items were found to have DIF. Country differences were examined to identify the possible cause of DIF and the results were found to be significant.

The examination of the measurement invariance of the data obtained from the American and Turkish students who speak different languages revealed that the steps up to the scalar invariance in the MG-CFA were established. This result parallels with the measurement invariance results for the "Support for Scientific Inquiry" questionnaire for students, which was administered within the scope of PISA 2006, in the study conducted by Asil and Gelbal (2012). In the analyses, it was found out that none of the items disturbed the invariance in samples of countries having a similar culture (Australia-New Zealand); that two items disturbed the invariance in the samples of the countries speaking the same language but having different

cultures (Australia-USA); and nine items disturbed the invariance in the samples of countries having different languages and cultures (Australia-Turkey). When MG-LCA was used to examine the measurement invariance of data obtained from American and Turkish students who speak different languages, the measurement invariance remained in the heterogeneity step. This step corresponds to the configural invariance in MG-CFA. In the MRM, on the other hand, the measurement invariance for the USA-Turkey samples could not be established and some items were found to have DIF. The country variable was examined to find out the possible causes of DIF and country difference was found to be a possible cause. In the study, Yandı, Köse, Uysal and Oğul (2017) obtained similar results and found that the measurement invariance could not be established when the countries with different cultures as well as different languages were compared. Köse (2015) also came up with a similar result. According to the results obtained from the study, while the individual parameter estimates obtained by MRM were good in heterogeneous data sets, it was observed that MRM was not successful in determining the reason for the difference in item function in data sets with multi-category and small sample. Sırgancı (2019) examined the effect of the covariant (common) variable in determining the changing item function with the Mixed Rasch Model. According to the results obtained from the study, MRM's latent DMF determination power and correct decision percentage increased significantly when the covariant variable was included in the model.

In conclusion, MG-LCA can be claimed to be a good alternative to MG-CFA in cases where the data structure is continuous. The differences detected between MG-CFA and MG-LCA are also similar to the results of the study carried out by Yandı, Köse and Uysal (2017). Moreover, the results obtained from this study coincide with the results of the studies conducted by Kankaras, Vermunt and Moors (2011) in which the methods based on IRT, SEM and LCA were compared. The advantage of the Mixed Rasch Model is that it allows not only detecting the DIF but also interpreting its possible cause more directly. Thus, unlike MG-CFA, MRM provides very detailed information for item response profiles. Therefore, it was found that MRM would be helpful especially in examining the invariance of the measurement instruments if used in combination with MG-CFA (Quandt, 2011).

According to the results obtained from this study, first of all, it is recommended to test the invariance of the structures to be compared in the comparison studies of the countries participating in large-scale exams. In this study, methods with different theoretical foundations were used to test the measurement invariance at the scale and item level. Future studies can test the measurement invariance with IRT-based methods in addition to these methods.

There are studies in the literature testing the measurement invariance (Eid, Langeheine & Diener, 2003; Kankaras & Moors, 2010; Somer, Korkmaz, Sural & Can, 2009; Yandı, 2017; Yandı, Köse & Uysal, 2017). The common finding in these studies is that the measurement invariance results obtained by different methods differ from each other. Since each method has its own assumptions and statistical backgrounds and is based on its own data structure different results can be obtained. In conclusion, it is recommended to provide evidence for measurement invariance by means of different methods in future studies (Kankaras, Vermunt & Moors, 2011).

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 author(s).

Authorship contribution statement

Authors are expected to present author contributions statement to their manuscript such as;

Zafer Erturk: Investigation, %60, Methodology, %65, Resources, %50, Visualization, %60, Software, %55, Formal Analysis, %60, and Writing, %50, Supervision, **Esra Oyar**: Investigation, %40, Methodology, %35, Resources, %50, Visualization, %40, Software, %45, Formal Analysis, %40, and Writing, %50.

ORCID

Zafer ERTÜRK https://orcid.org/ 0000-0003-3651-7602
Esra OYAR https://orcid.org/ 0000-0002-4337-7815

5. REFERENCES

Anderson, J. C., & Gerbing, D.W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*(2), 155-173. https://doi.org/10.1007/BF02294170

Arim, R. G., & Ercikan, K. (2014). Comparability between the American and Turkish versions of the TIMSS mathematics test results. *Education & Science, 39*(172), 33-48.

Aryadoust, V. (2015). Fitting a mixture Rasch model to English as a foreign language listening tests: The role of cognitive and background variables in explaining latent differential item functioning. *International Journal of Testing, 15*(3), 216-238. https://doi.org/10.1080/15305058.2015.1004409

Aryadoust, V., & Zhang, L. (2016). Fitting the mixed rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing, 33*(4), 529-553. https://doi.org/10.1177/0265532215594640

Asil, M., & Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği [Cross-cultural equivalence of the PISA student questionnaire]. *Eğitim ve Bilim, 37*(166), 236-249.

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed rasch model to a reading comprehension test: Identifying reader types. Practical Assessment, Research & Evaluation, 18. 1-13. https://doi.org/10.7275/n191-pt86

Bahadır, E. (2012). *Uluslararası Öğrenci Değerlendirme Programı'na (PISA 2009) göre Türkiye'deki öğrencilerin okuma becerilerini etkileyen değişkenlerin bölgelere göre incelenmesi* [According Programme for International Student Assessment (PISA 2009), investigation of variables that affect Turkish students' reading skills by regions]. Unpublished master thesis, Hacettepe University, Institutes of Social Sciences, Ankara.

Başusta, N. B., & Gelbal, S. (2015). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği [Examination of measurement invariance at groups' comparisons: a study on PISA student questionnaire]. *Hacetepe Üniversitesi Eğitim Fakültesi Dergisi, 30*(4), 80-90.

Bilir, M. K. (2009). *Mixture item response theory-mimic model: simultaneous estimation of differential item functioning for manifest groups and latent classes* (Unpublished doctoral dissertation). Florida State University.

Bowden, S. C., Saklofske, D. H., & Weiss, L. G. (2011). Invariance of the measurement model underlying the Wechsler Adult Intelligence Scale-IV in the United States and Canada. *Educational and Psychological Measurement, 71*(1), 186-199.

Brien, M., Forest, J., Mageau, G. A., Boudrias, J. S., Desrumaux, P., Brunet, L., & Morin, E. M. (2012). The basic psychological needs at work scale: measurement invariance between Canada and France. *Applied Psychology: Health and Well-Being, 4*(2), 167-187.

Bryne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34*(2), 155-175. https://doi.org/10.1177/0022022102250225

Buchholz, J., & Hartig, J. (2017). Comparing attitudes across groups: An IRT-based item-fit statistic for the analysis of measurement invariance. *Applied Psychological Measurement, 43*(3), 241-250. https://doi.org/10.1177/0146621617748323

Büyüköztürk, Ş. (2017). *Sosyal bilimler için veri analizi el kitabi istatistik, araştırma deseni SPSS uygulamaları ve yorum*. [Data analysis handbook statistics for social sciences, research design spss applications and interpretation.] Ankara: Pegem Akademi Yayıncılık.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. https://doi.org/10.1037/0033-2909.105.3.456

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-orderstructure: A walk through the process. *Structural Equation Modeling, 13*(2), 287-321. https://doi.org/10.1207/s15328007sem1302_

Cho, S. J. (2007). *A multilevel mixture IRT model for DIF analysis* (Doctoral dissertation, uga).

Cohen, A.S., & Bolt, D.M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*(2), 133–148. https://doi.org/10.1111/j.1745-3984.2005.00007

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55*(4), 1019-1031. https://doi.org/10.1037/0022-006X.55.4.584

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent trasition analysis: With applications in the social, behavioral, and health sciences*. New Jersey: John Wiley & Sons, Inc.

Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis: A primer. *Journal of Cross-Cultural Psychology, 34*(2), 195-210. https://doi.org/10.1177/0022022102250427

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

Fischer, G. H., & Molenaar, I. W. (Eds.). (2012). *Rasch models: Foundations, recent developments, and applications.* New York: Springer Science & Business Media.

Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement, 75*(2), 208-234. https://doi.org/10.1177/0013164414536183.

Goodman L. (2002) Latent class analysis In, Hagenaars J., McCutcheon A. (Ed.), *Applied latent class analysis (pp. 3-18)*. Cambridge University Press: New York.

Gülleroğlu, H. D. (2017). PISA 2012 matematik uygulamasına katılan Türk öğrencilerin duyuşsal özeliklerinin cinsiyete göre ölçme değişmezliğinin incelenmesi. [An investigation of measurement invariance by gender for the turkish students' affective characteristics whotook the PISA 2012 math test]. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 37*(1), 151-175.

Güngör, D., Korkmaz, M., & Somer, O. (2013). Çoklu-grup örtük sınıf analizi ve ölçme eşdeğerliği. [Multi-Group Latent Class Analysis and Measurement Equivalence]. *Türk Psikoloji Dergisi, 28*(72), 48-57.

Güzeller, O.C. (2011). PISA 2009 Türkiye örnekleminde öğrencilerin bilgisayar özyeterlik inançları ve bilgisayar tutumları arasındaki ilişkinin incelenmesi. [An ınvestigation of the relationship between students' computer self-efficacy beliefs and their computer attitudes in PISA 2009 turkey sampling] *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 12*(4), 183-203.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *The Electronic Journal of Business Research Methods. 6*(1) 53 – 60.

Horn, J. L., McArdle, J.J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *The Southern Psychologist, 1*(4), 179-188.

Hui, C.H., & Triandis, H.C. (1985). Measurement in cross-cultural psychology: a review and comparison of strategies. *Journal of Cross-cultural Psychology, 16*(2), 131–152. https://doi.org/10.1177/0022002185016002001

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.

Kankaras, M. (2010). *Essays on measurement equivalence in cross-cultural survey research: A latent class approach* (Unpublished doctoral dissertation).

Kankaras, M., & Moors, G. (2010). Researching measurement equivalence in cross-cultural studies. *Serbian Psychological Association, 43*(2), 121-136.

Kankaras, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, 40(2), 279-310. https://doi.org/10.1177/0049124111405301

Karakoc Alatli, B., Ayan, C., Polat Demir, B., & Uzun, G. (2016). Examination of the TIMSS 2011 Fourth Grade Mathematics Test in terms of cross-cultural measurement invariance. *Eurasian Journal of Educational Research, 66*, 389-406. https://doi.org/10.14689/ejer.2016.66.22

Karasar, N. (2013). *Bilimsel araştırma yöntemi*. [Scientific research methods]. Ankara: Nobel Yayınevi.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27*(4), 307-327. https://doi.org/10.1111/j.1745-3984.1990.tb00751.x

Köse, İ. A. (2015). PISA 2009 öğrenci anketi alt ölçeklerinde (q32-q33) bulunan maddelerin değişen madde fonksiyonu açısından incelenmesi. [Examining the differential item functioning in the PISA 2009 student survey subscales (q32-q33)] *Kastamonu Eğitim Dergisi, 23*(1), 227-240.

Langeheine, R., Pannekoek, J., & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research, 24*(4), 492-516. https://doi.org/10.1177/0049124196024004004

Li, M., Wang, M. C., Shou, Y., Zhong, C., Ren, F., Zhang, X., & Yang, W. (2018). Psychometric properties and measurement invariance of the brief symptom inventory-18 among chinese insurance employees. *Frontiers in psychology, 9*, 519. https://doi.org/10.3389/fpsyg.2018.00519

Liang, L., & Lee, Y. H. (2019). Factor structure of the ruminative response scale and measurement ınvariance across gender and age among chinese adolescents. *Advances in Applied Sociology, 9*, 193-207. https://doi.org/10.4236/aasoci.2019.96016

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological methodology, 31*(1), 223-264. https://doi.org/10.1111/0081-1750.00096

McCutcheon, A. L., & Hagenaars, J. A. (1997). Comparative social research with multi-sample latent class models. *Applications of latent trait and latent class models in the social sciences*, 266-277.

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

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515. https://doi.org/10.1207/S15327906MBR3903_4

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195-215. https://doi.org/10.1007/BF02295283

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities. A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review, 20*(4), 303-320. https://doi.org/10.1093/esr/jch026

Moors, G., & Wennekers, C. (2003). Comparing moral values in Western European countries between 1981 and 1999. A multiple group latent-class factor approach. *International Journal of Comparative Sociology, 44*(2), 155-172. https://doi.org/10.1177/002071520304400203

Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies, 26*(3), 573-596. https://doi.org/10.1057/palgrave.jibs.8490187

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Oon Pey Tee., & R. Subramaniam (2018) Comparative study of middle school students' attitudes towards science: Rasch analysis of entire TIMSS 2011 attitudinal data for England, Singapore and the U.S.A. as well as psychometric properties of attitudes scale. *International Journal of Science Education, 40*(3), 268-290. https://doi.org/10.1080/09500693.2017.1413717

Ölçüoğlu, R. (2015). *TIMMS 2011 Türkiye sekizinci sınıf matematik başarısını etkileyen değişkenlerin bölgelere göre incelenmesi* [The investigation of the variables that affecting TIMSS 2011 Turkey eight grade math achievement according to regions]. Unpublished master thesis, Hacettepe University, Institutes of Social Sciences, Ankara.

Ölmez, İ. B., & Cohen, A. S. (2018). A mixture partial credit analysis of math anxiety. *International Journal of Assessment Tools in Education, 5*(4), 611-630. https://doi.org/10.21449/ijate.455175

Önen, E. (2009). *Ölçme değişmezliğinin yapısal eşitlik modelleme teknikleri ile incelenmesi* [Examination of measurement invariance with structural equation modelling techniques]. Unpublished doctoral thesis, Ankara University, Ankara.

Pishghadam, R., Baghaei, P., & Seyednozadi, Z. (2017). Introducing emotioncy as a potential source of test bias: A mixed Rasch modeling study. *International Journal of Testing, 17*(2), 127-140. https://doi.org/10.1080/15305058.2016.1183208

Quandt, M. (2011). Using the mixed Rasch model in the comparative analysis of attitudes. *Cross-cultural analysis: Methods and applications,* 433-460.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*(1), 75-92. https://doi.org/10.1111/j.2044-8317.1991.tb00951.x

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

Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. *In Rasch models (pp. 257-268)*. Springer: New York, NY.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31-57. https://doi.org/10.1177/0013164413498257

Schnabel, D. B., Kelava, A., Van de Vijver, F. J., & Seifert, L. (2015). Examining psychometric properties, measurement invariance, and construct validity of a short version of the Test to Measure Intercultural Competence (TMIC-S) in Germany and Brazil. International *Journal of Intercultural Relations, 49,* 137-155. https://doi.org/10.1016/j.ijintrel.2015.08.002

Sırgancı, G. (2019). *Karma rasch model ile değişen madde fonksiyonunun belirlenmesinde kovaryant (ortak) değişkenin etkisi*. [The effect of covariant (common) variable in determining the changing item function with mixed rasch model]. Unpublished doctoral thesis, Ankara University, Faculty of Education, Ankara.

Silvia, P. J., Kaufman, J. C., & Pretz, J. E. (2009). Is creativity domain-specific? Latent class models of creative accomplishments and creative self-descriptions. *Psychology of Aesthetics, Creativity, and the Arts, 3*(3), 139-148. https://doi.org/10.1037/a0014940

Somer, O., Korkmaz, M., Dural, S., & Can, S. (2009). Ölçme eşdeğerliliğinin yapısal eşitlik modellemesi ve madde tepki kuramı kapsamında incelenmesi. [Examining measurement invariance with structural equation modeling and item response theory]. *Türk Psikoloji Dergisi, 24*(64), 61-75. https://doi.org/10.14527/9786053188407.23

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-90. https://doi.org/10.1086/209528

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar. [Structural equation modeling: basic concepts and lisrel applications]. *Türk Psikoloji Yazıları, 3*(6), 49-74.

Şen, S. (2016). Applying the mixed Rasch model to the Runco ideational behavior scale. *Creativity Research Journal, 28*(4), 426-434. https://doi.org/10.1080/10400419.2016.1299858

TIMSS (2011). TIMSS 2011 international database. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat, Amsterdam, the Netherlands. Retrieved January 10, 2020 from http://timss.bc.edu/timss2011/international-database.html.

Tucker, K. L., Ozer, D. J., Lyubomirsk, S., & Boehm, J. K. (2006). Testing for measurement invariance in the satisfaction with life scale: A comparison of Russians and North Americans. *Social Indicators Research, 78*(2), 341–360. https://doi.org/10.1007/s11205-005-1037-5

Uyar, Ş. & Doğan, N. (2014). PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi [An investigation of measurement invariance of learning strategies model across different groups in PISA Turkey sample]. *Uluslararası Türk Eğitim Bilimleri Dergisi,* 2(3), 30-43

Uzun, N. B. (2008). *TIMSS-R Türkiye örnekleminde fen başarısını etkileyen değişkenlerin cinsiyetler arası değişmezliğinin değerlendirilmesi* [Assessing the measurement invariance of factors that are related to students' science achievement across gender in TIMSS-R Turkey sample]. Unpublished master thesis, Hacettepe University, Institutes of Social Sciences, Ankara.

Uzun, B. & Ogretmen T. (2010). Fen başarısı ile ilgili bazı değişkenlerin TIMSS-R Türkiye örnekleminde cinsiyete göre ölçme değişmezliğinin değerlendirilmesi. [Assessing the measurement invariance of factors that are related to students' science achievement across gender in TIMSS-R Turkey Sample]. *Eğitim ve Bilim, 35*(155), 26-35.

von Davier M. (2001). WINMIRA 2001: Software and user manual. Available from: http://208.76.80.46/~svfklumu/wmira/ index.html.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. *In Rasch models (pp. 371-379)*. Springer, New York, NY.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*(3), 1-26. https://doi.org/10.7275/mhqa-cd89.

Yalçın, S. (2019). Use of mixed item response theory in rating scales. *International Electronic Journal of Elementary Education, 11*(3), 273-278.

Yandı, A. (2017). *Ölçme eşdeğerliğini incelemede kullanılan yöntemlerin farklı koşullar altında istatistiksel güç oranları açısından karşılaştırılması* [Comparison of the methods of examining measurement equivalence under different conditions in terms of statistical power ratios]. Unpublished doctoral thesis, Ankara University, Institutes of Social Sciences, Ankara.

Yandı, A., Köse, İ. A., & Uysal, Ö. (2017). Farklı yöntemlerle ölçme değişmezliğinin incelenmesi: Pisa 2012 örneği. [Examining the measurement invariance with different methods: Example of Pısa 2012] *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 13*(1), 243-253. https://doi.org/10.17860/mersinefd.305952

Yandı, A., Köse, İ. A., Uysal, Ö., & Oğul, G. (2017). *PISA 2015 öğrenci anketinin (st094q01nast094q05na) ölçme değişmezliğinin farklı yöntemlerle incelenmesi.* [Investigation of the PISA 2015 student survey (ST094Q01NA-ST094Q05NA) with the different methods of measurement]. Ankara: Pegem

Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika, 62*(3), 297-330. https://doi.org/10.1007/BF02294554

Yüksel, S. (2015). *Ölçeklerde saptanan madde işlev farklılığının karma rasch modelleri ile incelenmesi* [Analyzing differential item functioning by mixed rasch models which stated in scales]. Unpublished doctoral thesis, Ankara University, Institutes of Health Sciences, Ankara.

Yüksel, S., Elhan, A. H., Gökmen, D., Küçükdeveci, A. A., & Kutlay, Ş. (2018). Analyzing differential item functioning of the Nottingham Health Profile by mixed rasch model. *Turkish Journal of Physical Medicine & Rehabilitation, 64*(4), 300-307. https://doi.org/10.5606/tftrd.2018.2796

6. APPENDIX

Table A1. *Items in the Mathematics Liking Scale*

Codes	Items - English	Items - Turkish
BSBM17A	I enjoy learning mathematics	Matematik öğrenirken eğleniyorum.
BSBM17B	I wish I did not have to study mathematics*	Keşke matematik çalışmak zorunda olmasam.*
BSBM17C	Mathematics is boring*	Matematik sıkıcıdır.*
BSBM17D	I learn many interesting things in mathematics	Matematik dersinde ilginç şeyler öğrenirim.
BSBM17E	I like mathematics	Matematiği severim.
BSBM17F	I like any schoolwork that involves numbers	Sayıların dâhil olduğu her okul işini severim.
BSBM17G	I like to solve mathematics problems	Matematik problemlerini severim.
BSBM17H	I look forward to mathematics class	Matematik derslerini dört gözle beklerim.
BSBM17I	Mathematics is one of my favorite subjects	Matematik favori dersimdir.

*Reverse scored items (TIMSS, 2015).