ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 120-137

Comparing Differential Item Functioning Based on Multilevel Mixture Item Response Theory, Mixture Item Response Theory and Manifest Groups

Ömer DOĞAN* Burcu ATAR**

Abstract

Studies on the differential item functioning (DIF) are usually considered in the context of manifest groups. Recently, with the increase in the number of analyses conducted with mixture models, investigating the situations that cause differences between groups has come to the forefront. In addition, it is considered important to examine the DIF with mixture models in which levels are also handled. In this study, it is aimed to compare the results of the multilevel mixture item response theory (MMIRT) model and the mixture item response theory (MIRT) model and the results of the DIF analyses based on the manifest groups. The research sample consists of students who answered the second booklet in the electronic Trends in International Mathematics and Science Study (eTIMSS) 2019 and coded their gender. The answers given to 15 items were analyzed with the Mantel Haenszel (MH) method for the gender variable according to the manifest groups, and with the selection of the most appropriate models by varying the number of groups and the number of levels according to the MIRT model and the MMIRT model. DIF analyses of the obtained latent groups were also performed with the MH method. In the light of the findings, the number of items displaying DIF in both the MIRT model and the MMIRT model is higher than the manifest groups. While only one item displayed DIF in the analysis according to gender, 14 items displayed DIF according to the MIRT model and seven items displayed DIF according to the MMIRT model. There is not a complete overlap in the number of DIF items and DIF effect sizes found as a result of the MIRT model and MMIRT model analyses. For this reason, a level analysis should be conducted before the analyses and if there is multi-levelness, the analyses should be conducted by taking this situation into consideration.

Keywords: multilevel mixture item response theory model, mixture item response theory model, manifest groups

Introduction

In education, various tests are applied to determine the level of acquisition of the skills desired to be gained by individuals, to identify learning deficiencies and to place individuals in various institutions. In order to prevent errors in the tasks to be carried out through the scores obtained from these tests, several precautions are taken within the scope of measurement and evaluation. The fact that the scores of a test are valid and reliable contributes to the fairness of the decisions to be made using the scores. Validity, which is the first of these two important concepts, also includes reliability. Validity is a concept whose definition and content are constantly renewed according to the point of view in the historical process. Standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) define validity as the degree to which interpretations of test scores are supported by evidence and theory. Accordingly, validity is not a characteristic of the test, but is related to the inferences made from the test scores. The validity process also involves gathering the necessary evidence for a sound scientific basis for the proposed score interpretations. One of the evidences that should be obtained in this process can be obtained by analyzing differential item functioning (DIF), which is one of the evidences about the internal structure of the test. According to Kelderman and Macready (1990), test items exhibit DIF if the item scores of equal ability

* PhD Student., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: 64omerdogan64@gmail.com, ORCID ID: 0000-0001-5169-520X
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
Doğan, Ö., & Atar, B. (2024). Comparing differential item functioning based on multilevel mixture item response theory, mixture item response theory and manifest groups. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(2), 120-137. https://doi.org/10.21031/epod.1457880

Received: 24.03.2024
Accepted: 6.06.2024

test takers from different groups (e.g., different gender, race, region or age) are significantly different. If a number of items on a test display DIF in favor of a specific group, there may be an unfair advantage for that group in terms of the assessed level of performance when compared to individuals from other groups. Items of test that display DIF are one of the important reasons for reducing the validity of the scores (Kristanjansonn et al., 2005; Messcik, 1995). DIF is an important indicator of test quality because it is directly related to the fairness and validity of the test. There are many methods for determining DIF, including Mantel-Haenszel (Holland & Thayer, 1988; Mantel & Haenszel, 1959), logistic regression, analysis of variance, transformed item difficulty and SIBTEST (Shealy & Stout, 1993) within the framework of classical test theory (CTT), Lord's (1980) chi-square method, Raju's (1988, 1990) field measurements and likelihood ratio test (Thissen et al.,1988) within the framework of item response theory (IRT). In DIF detection, the above-mentioned methods are compared with groups that are considered to be homogeneous within themselves, namely focal and reference groups. These groups are formed by gender, ethnicity, nationality, etc. and are referred to as manifest groups.

DIF detection methods in the context of CTT and IRT are very useful for detecting DIF in test administration, but they have made little progress in understanding the possible causes of DIF. This is because manifest group characteristics are typically only marginally related to the cause of DIF (Choi et al., 2015; Roussos & Stout, 1996). Several studies have shown that the homogeneity assumption is not always met in DIF analysis of manifest groups (e.g., Cohen & Bolt, 2005; de Ayala et al., 2002). Moreover, when differences between groups are found, it is not easily understood who is primarily advantaged or disadvantaged by DIF items (de Ayala et al., 2002).

Methods for DIF detection that have been mentioned in the context of IRT include comparisons of item parameters or areas between item response functions. However, efforts to understand why some test takers respond differently to these items are often conducted outside of the IRT context. Mixture IRT (MIRT) models have been proposed as a useful tool to investigate how differences in qualitative test takers, such as differences resulting from the use of different problem solving strategies, can lead to differences in responses to test items (Embretson & Reise, 2000). The use of the MIRT model, which is an integration of the IRT and latent class models, is typically exemplified by comparisons of item profiles across different latent groups or latent classes (Paek & Cho, 2015).

MIRT model is similar to a multigroup item response model, but the group of interest is not predetermined, but is determined based on the results obtained from model parameter estimation. As in multigroup item response models, item parameters and latent variable(s) may be different across latent groups in MIRT models (Cho et al., 2015). In MIRT models, individuals are assigned to non-predetermined classes with the highest within-group homogeneity and highest between-group heterogeneity in terms of the latent trait. Item parameters are estimated independently of the manifest group to which the individuals belong and specific to each group. Differences in group-specific estimated parameters suggest that DIF may be caused by a latent trait (De Ayala et al., 2002). De Boeck et al. (2011, p. 584) list four a priori reasons to consider implicit DIF analysis instead of manifest DIF analysis:

1. Lack of opinion (no idea about which group membership is interesting, or incomplete knowledge of group membership),

2. Unobservability (the group membership of interest is not observable),

3. Reliability (observed group membership may not be completely reliable) and

4. Validity (observed group membership may not be a completely valid indicator of actual group membership).

In the context of DIF models, Cohen and Bolt (2005) described a mixture Rasch model (MRM) approach to detecting uniform DIF, which differs from previous methods in some fundamental respects. This MRM is expressed as follows:

$$P(y_{ij} = 1|\theta_i) = \sum_{g=1}^{G} \pi_g \frac{\exp[(\theta_{ig} - b_{ig})]}{1+\exp[(\theta_{ig} - b_{ig})]} \quad (1)$$

g= 1,.....,G: Index indicating the latent class

j= 1,.....,J: Index indicating respondents

$\theta jg = j$. Individual's latent ability in latent class g

β= item difficulty parameter of item i in class g

Besides the MRM, there are also 2-parameter and 3-parameter models for mixture models. The two-parameter Mixture IRT model is shown as follows (Finch & French, 2012):

$$P(y_{ij} = 1|\theta_i) = \sum_{g=1}^{G} \pi_g \frac{\exp\left[a_{jg}\left(\theta_{ig} - b_{ig}\right)\right]}{1+\exp\left[a_{jg}\left(\theta_{ig} - b_{ig}\right)\right]} \quad (2)$$

The three-parameter Mixture IRT model, which includes item parameters and chance parameter for each grade, is shown as follows (Choi et al., 2015):

$$P(y_{ij} = 1|\theta_i) = \sum_{g=1}^{G} \pi_g [c_{jg} + (1 - c_{jg}) \frac{\exp\left[a_{jg}\left(\theta_{ig} - b_{ig}\right)\right]}{1+\exp\left[a_{jg}\left(\theta_{ig} - b_{ig}\right)\right]} \quad (3)$$

It can be said that MIRT models are important factors in the estimation of item parameters. In their study, Cohen and Bolt (2005) used mixture models to decompose the secondary dimension expressed by Ackerman (1992) and aimed to better understand the differences between test takers who were disadvantaged or advantaged by DIF items. In Study 1, they showed that the conventional approach to studying DIF does not contribute much to understanding the causes of DIF. They concluded that using explicit gender categories to identify those affected by gender DIF is likely to be misleading. Study 2 extended the analysis of DIF, showing how mixture models can be used to identify latent groups where some form of DIF may be present in the first place. In the case of the groups in Study 2, it was explained that there is a cognitive interpretation of the secondary dimension and thus the cause of the DIF can be more easily interpreted. As a result, in the case of gender DIF, it was clear that not all members of a gender group responded in the same way to items that were allegedly biased for or against their group, with some men being disadvantaged by items that were found to advantage men and some women being advantaged by items that were found to disadvantage women. Therefore, when it is accepted that DIF items do not universally advantage or disadvantage all members of a group, this practice becomes questionable. Similarly, Samuelsen (2005) based the basic premise of his study on the fact that it is not advisable to use open groups in DIF analyses. She argued that distinctions based on external characteristics of test takers are not useful and that the groups that emerge are neither homogeneous nor cognitively meaningful. Instead, by examining the latent dimensions underlying student performance, it is possible to identify and interpret the reasons behind DIF. By using the latent class perspective, individual differences in human behavior can be attributed to potentially meaningful dimensions rather than external characteristics, and when this happens, it is possible to truly explain why items work differently. In their study, Jiao and Chen (2014) addressed the problems arising from the use of the DIF approach based on traditional observed groups and analyzed both background and cognitive covariates that are effective in the characterization of latent class membership. The results of the study showed that a sole manifest group variable is insufficient to fully predict the sources of implicit DIF and that the implicit class-based DIF approach is a possible method for screening for potential DIF items arising from the intervening effects of multiple variables. The aforementioned studies and others (Cho & Cohen, 2010; Dras, 2023; Zhang, 2017) have shown that the MRM approach can provide more insight into the antecedents of DIF than methods that rely on assessing DIF in relation to manifest groups. In addition, this approach to DIF assessment has the potential to provide more comprehensive analyses that do not rely on a predetermined ranking of individuals, which itself may be biased in some respects (Finch & Finch, 2013). The mixture model is used to define latent classes of test takers who are homogeneous in terms of their item response patterns. Members of each latent class differ in ability and response strategies differ across classes. However, an important limitation of the mixture model is that it essentially ignores the underlying multilevel structure that exists beyond the student level in most educational test data (Cho & Cohen, 2010).

If the analysis is restricted to the traditional linear model, the basic assumptions are normality, homoscedasticity, linearity, and independence. It is desirable to preserve normality and linearity in the analyses, but the assumption of homoscedasticity and especially the assumption of independence need to be adapted. The general idea behind such adaptations is that persons in the same group are closer or more similar than persons in different groups. Thus, individuals in different classes may be independent, but individuals in the same class share values on many more variables (Raudenbush & Bryk, 2002). The biggest threat to the local independence assumption is the nested data structure (Jiao et al., 2012). For example, the multilevel data structure manifest in achievement tests is a structure in which students are nested to teachers and teachers are nested to schools. In addition to the mixture model, a fairly recent contribution to the DIF literature has been the emergence of methods for dealing with the multilevel data structure that is common in such assessments (French & Finch, 2010). For instance, especially in large-scale assessments, data for DIF detection studies are often collected from test takers nested within schools. In such cases, schools should be assumed to influence test item responses, at least to some extent. This influence will be expressed in the form of non-trivial intracluster correlation (ICC) values. When such multilevel data structure is ignored and ICCs are non-zero (or very close to it), the resulting analyses are likely to yield erroneous estimates of item parameters and their associated standard errors, leading to erroneous DIF detection results. Researchers (e.g. Finch &French, 2012) have continued to develop and adapt multilevel methods for DIF detection in the context of manifest groups (Finch & Finch, 2013).

Cho and Cohen (2010) described the MMIRT model, which allows for the simultaneous detection of differences in latent class structure at both test taker and school levels. Student-level latent classes capture the relationship between responses in the student-level unit. The MIRT model assumes that there may be heterogeneity in response patterns at the first level that should not be ignored (Mislevy & Verhelst, 1990; Rost, 1990). However, the MMIRT model also takes into account the possibility that there may not be latent classes at the first level. (Cho, 2007).

In the MMIRT model, dependency is taken into account by including latent variables at higher-level continuous and/or categorical latent variables. Vermunt (2007) proposed eight possible versions of two-level (e.g., students nested within schools) MMIRT models. Latent variables at each level of mixture models can be categorical, continuous, or both categorical and continuous, as mixture models include categorical latent variables and item response models include continuous latent variables (as cited in Lee et al., 2018).

Cho and Cohen (2010) showed in their study that it is possible to obtain grade-specific item difficulties for each level 1 and 2 and express them on the same scale. In the empirical example they examined, the mixed groups at the student and school level that emerged in the data were similarly clearly distinguishable in terms of ability levels, item difficulty profiles, student and school demographics, and response patterns, but when more than one factor characterizes a class, it can be difficult to find factors that potentially cause DIF. Gurkan (2021) used Programme for International Student Assessment (PISA) 2012 data to investigate the correlation patterns of the multidimensional and multilevel MIRT model and to improve the model, and aimed to investigate the variance between within-country correlations based on traditional estimates and to determine to what extent this variance is due to heterogeneity in the amount of measurement error and the clustered nature of the data. As required by the characteristics of the PISA data, the multidimensional MMIRT models used in the study not only appropriately accounted for measurement error and clustering in the data, but also took into account the possibility of different subpopulations within countries.

Another international study of the PISA type is TIMSS. TIMSS is an international comparative study that measures student achievement in mathematics and science worldwide. Conducted in a four-year assessment cycle since 1995, TIMSS has assessed student achievement in fourth and eighth grades seven times - 1995, 1999, 2003, 2007, 2011, 2015 and 2019 - and accumulated 24 years of trend measurements. In 2019, TIMSS began transitioning to computer-based assessment by introducing a digital version of the paper-and-pencil assessment called “eTIMSS”. Within the scope of the research, the use of real data was planned and eTIMSS 2019 data was utilized. This is because the DIF studies conducted with MIRT

models, which have been increasing recently, are mostly conducted using simulative data (e.g. Cho, 2007; Cho & Cohen, 2010; de Ayala et al. 2002; Sırgancı, 2019; Uyar, 2015). In the current studies, deficiencies such as disregarding the levelness (Choi et al., 2015; Toker & Green, 2021; Yalcin, 2018, etc.), conducting DIF analysis based only on manifest groups (Aydemir, 2023; Bayram, 2024; Unal, 2023, etc.), lack of using real data (Sırgancı, 2019; Uyar, 2015, etc.) and ignoring the source of DIF were found. The aim of this study is to compare the results of DIF analyses based on the MMIRT model and manifest groups and to investigate what may cause performance differences in eTIMSS. In the study, DIF analysis was performed on the data in eTIMSS booklet 2 with the MMIRT model and the results obtained were compared with the results of DIF analysis based on the MMIRT model and manifest groups. In the light of the findings obtained, the number of latent classes, items with DIFs and changes in the number of items with DIFs were examined when multi-levelness and the differentiation of manifest groups and latent groups were included in the analyses in studies such as TIMSS prepared for cross-country comparisons in education. Thus, by comparing mixture models and manifest groups methods, the differences in determining the source of DIF were revealed and it was investigated whether the addition of multi-levelness to the mixture model had a positive effect on the complexity of the model.

Methods

Sample

In this study, the typical case sampling method of purposive sampling was used. Since the models used in the DIF analysis (MMIRT and MIRT) are based on the IRTTIMSS items developed according to this model were used. In 2019, TIMSS started to move to computer-based assessment by introducing a digital version of the paper-and-pencil assessment called "eTIMSS". This included 22 countries at the eighth grade level and five participants from regions or cities of some countries as benchmark participants.

In the study, the second booklet was selected because it is suitable for multilevel data structure and the number of multiple-choice items is higher than the other booklets. For the study, the responses of eighth grade students from 22 countries in the eTIMSS 2019 data to 15 dichotomously scored mathematics and science items in the second booklet were used. Within the scope of the research, the answers of 8167 individuals were analyzed and 123 individuals were excluded from the study because their gender was not specified. Finally, the data of 8044 individuals were analyzed. Li et al., (2009) stated that a sample size of 600 individuals would be appropriate for MIRT models when the number of items is between 15 and 30. In addition, Li et al. (2009) stated that for a 15-item test, a sample size of 600 would be sufficient in a model with 1 to 4 classes for both MIRT 2PL and MIRT 3PL models. Cho et al., (2013) state that a sample size of more than 360 can be used for the MRM. Cohen and Bolt (2005) successfully applied the MIRT 3PL model with a sample size of 1000. Demographic information is presented in Table 1.

Table 1

22 eTIMSS participant countries, number of participants and average scores

Country	Number of Participants	Mean Score	Gender(F/M)
United Arab Emirates (UAE)	1584	6.42	774/810
Chile	289	5.42	141/148
England	222	7.00	120/102
Finland	347	7.27	178/169
France	266	5.37	139/127
Georgia	244	5.82	118/126
Hong Kong	228	8.64	109/119
Hungary	328	7.65	181/147
Israel	267	7.36	141/126
Italy	257	6.12	132/125
Korea Rep. of	273	10.79	137/136

Table 1 (continued)

Country	Number of Participants	Mean Score	Gender(F/M)
Lithuania	259	6.95	125/134
Malaysia	499	7.50	258/241
Norway	335	6.96	156/159
Portugal	238	6.45	122/116
Qatar	278	6.06	129/149
Russian Federation	278	8.33	125/153
Singapore	352	10.53	178/174
Sweden	280	7.38	127/153
Türkiye	289	7.11	137/152
Chinese Taipei	349	10.58	178/171
United States	602	7.55	278/324
General	8044	7.33	3983/4061

As seen in Table 1, the number of male and female students is close to each other. The highest number of participants was from the UAE, while the lowest number of participants was from England. Looking at the mean scores, the three highest scores belong to the states located in Asia.

Data Analysis

The eTIMSS 2019 application consists of 14 booklets. The booklets contain mathematics and science items with certain common items. The items are prepared as multiple-choice, open-ended and short-answer. Within the scope of the study, 31 items from mathematics and science courses were selected, all of which were four-choice multiple-choice items. ICC values and dimensionality structure of the items were examined for the planned MMIRT model. Students were identified as level 1 and countries as level 2. The fact that the ICC values are close to zero indicates that there is no nested structure. For this reason, items with ICC values close to zero were removed and the analyses continued with the remaining 15 items. The average of the ICC values of the selected items is approximately .15. In other words, approximately 15% of the variance is due to country differences. Muthen (1997) suggested that multilevel modeling should definitely be taken into account when group sizes exceed 15 if the ICC>.10, and Julian (2001) and Selig et al., (2008) suggested that the hierarchical structure should not be ignored even when the ICC values are lower than .10 (as cited in Şen, 2022). The ICC values for 15 items are given in Table 2.

Table 2.

ICC values for the selected 15 items

Item Numbers	ICC Values
1	.12
2	.11
3	.17
4	.15
5	.14
6	.13
7	.17
8	.16
9	.20
10	.10
11	.11
12	.12
13	.17
14	.15
15	.19
Mean	.15

According to Table 2, the lowest ICC value is .10 while the highest value is .19 These values indicate that at least 10% of the variance of each item is due to country differences. Regarding the 15 items used in the study, it was examined whether there was a unidimensional structure. In order to determine this, the suitability of the data for factor analysis was examined using the 'fa' function in the 'psych' package of R software (Revelle, 2023) and exploratory factor analysis (EFA) based on the tetrachoric correlation matrix was performed on the data. The adequacy of the correlation matrix between the items and its comparison with the unit matrix were examined with Kaiser-Meyer-Olkin (KMO) coefficient and Barlett's test of sphericity. For factorization, the KMO value is expected to be higher than 0.60 and the Barlett test is expected to be significant (Büyüköztürk, 2018). For 15 items, the KMO value was found to be 0.850 and the Barlett test was significant ($p<.001$). Therefore, it was interpreted that the data was appropriate for factorization. The eigenvalues obtained in the analyses for dimensionality are shown in Table 3 and the slope accumulation graph is shown in Figure 1. According to the values obtained, it is understood that the data shows a unidimensional structure.

Table 3.

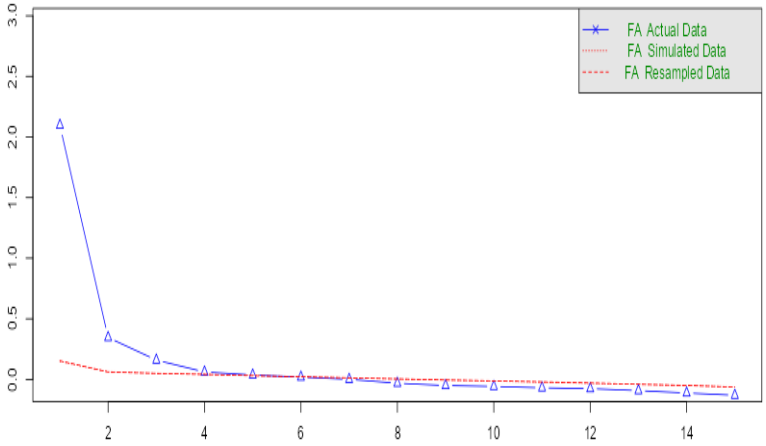
Eigenvalues obtained in dimensionality analyses for 15 items

Factor Number	Eigenvalue
1	2.10
2	.35
3	.16
4	.06
5	.04

According to Table 3, only the eigenvalue for the first dimension is greater than 1, the others are less than 1 and the ratio between the first two eigenvalues is six times. According to Kaiser's (1960) K1 rule, the construct is unidimensional.

Figure 1.

Slope deposition graph for 15 items



The courses, subject areas and cognitive domains of the 15 items selected for the analysis are given in Table 4. Accordingly, the subject areas of the items selected from seven mathematics and eight science courses consist of numbers, algebra, geometry, biology, chemistry, earth science, and physics. In addition, there are items from all three cognitive domains of eTIMSS: knowing, applying, and reasoning.

Table 4.

Courses, subject areas and cognitive domains of the data

Item Number	Course	Subject Area	Cognitive Domaim
1	Math	Numbers	Applying
2	Math	Algebra	Knowing
3	Math	Algebra	Knowing
4	Math	Algebra	Knowing
5	Math	Algebra	Knowing
6	Math	Geometry	Applying
7	Math	Geometry	Reasoning
8	Science	Biology	Knowing
9	Science	Chemistry	Knowing
10	Science	Earth Science	Knowing
11	Science	Earth Science	Reasoning
12	Science	Earth Science	Applying
13	Science	Chemistry	Applying
14	Science	Biology	Applying
15	Science	Physics	Knowing

According to Table 4 in the math section includes one item on numbers, four items on algebra and two items on geometry. In the science section, there are two items each from biology and chemistry, one item from physics and three items from earth science. Three information criterion indices are used to determine the appropriate model for parameter estimation based on the MIRT and MMIRT models. Akaike's (1974) information criteria (AIC), Schwarz's (1978) Bayesian information criterion (BIC) and the sample size-adjusted version of BIC (SABIC; Sclove, 1987). Within the scope of the research, BIC value is used in accordance with the literature (Choi et al., 2015; Li et al. 2009; Şen & Toker, 2021).

The Mplus software package was used to determine the appropriate model based on the DIF according to the MIRT and MMIRT (Muthén & Muthén, 2017). The robust version of the marginal maximum likelihood estimation technique (MLR) was used in Mplus parameter estimation. In addition, the number of iterations was increased in the analyses as the models became more complex. For the DIF analysis for manifest groups, the MH technique was chosen and the "difR" package in the R software language was used (Magis et al., 2015). Latent classes were characterized in terms of item difficulty parameter estimates and descriptive characteristics of test takers. As suggested by Cho and Cohen (2010), test taker-level DIF analyses were conducted separately for each second-level latent class, while uniform country-level DIF was determined by comparing school latent class item difficulty estimates across test taker levels. It was decided to use the standardized MH test when there were two latent classes and the Generalized Mantel Haenszel (GMH) when there were more latent groups. If $\Delta MH>0$, DIF is interpreted as DIF in favor of the focus group, $\Delta MH<0$ as DIF in favor of the reference group, and $\Delta MH\cong0$ as no DIF (Holland & Thayer, 1986).

Results

Within the scope of the study, 9 different models were analyzed for MIRT and MMIRT with the data set consisting of 15 items in eTIMSS booklet 2. Model fit statistics for these nine models are presented in Table 5.

Table 5

Model fit statistics for 9 models

Model	LogL	np	AIC	BIC	SABIC
L0-G2	-73891.28	61	147904.55	**148331.11**	148137.26
L0-G3	-73764.92	91	147711.84	148348.17	14805899
L0-G4	-73764.96	92	147713.91	148357.24	148064.88
L1-G2	-72759.79	91	145701.58	146337.92	146048.75
L1-G3	-72492.14	137	145258.28	146216.28	145780.92
L1-G4	-72372.38	183	145110.75	146390.41	145808.87
L2-G2	-71745.28	167	144138.24	**145306.02**	144775.33

Table 5 (continued)

Model	LogL	np	AIC	BIC	SABIC
L2-G3	-71591.89	243	143669.77	145368.99	144596.79
L2-G4	-71388.76	319	143415.56	145646.18	144632.46

LogL: Log-likelihood; np: Number of Parameter; AIC: Akaike's Information Criteria ; BIC: Bayesian Information Criterion; SABIC: Sample Size-Adjusted Version of BIC

As shown in Table 5, the level 0 and number of groups 2 (L0-G2) model has the smallest BIC value among the MIRT models. The level 2 and number of groups 2 (L2-G2) model has the smallest BIC value among the MMIRT models. As mentioned in the data analysis section, it is in the literature that BIC is more appropriate in the selection of mixture models. Therefore, in the light of these results, the L0-G2 model among the MIRT models and the L2-G2 model among the MMIRT models are used in the analyses. Based on the L0-G2 model, students are divided into two latent student classes, and based on the L2-G2 model into two latent student classes and two latent country classes. Table 6 and Table 7 present the final class numbers and proportions for each latent class variable based on the estimated posterior probabilities for the MIRT and MMIRT models. Student-level latent class 2 is the dominant group (.73) in the MIRT model. Note that the sum of the proportions reported in Table 6 is equal to 1. In the MMIRT model, the second level student level class 1 is the dominant group (.45).

Table 6.

Final Class Numbers and Ratios for Each Student Level Latent Classroom for the MIRT Model

Latent Class	Number of Individuals (Female/Male)	Ratio
1	2233(1149/1084)	.28
2	5811(2834/2977)	.72

Table 7.

Final Class Numbers and Ratios for the Student and Country Level Latent Class for the MMIRT Model

	Student Level Latent Group	
Country Level Latent Group	1	2
1	1499(741/758) (.19)	240(125/115) (.03)
2	5319(2640/2679) (.66)	986(477/509) (.12)

Table 6 and Table 7 present the final class numbers and proportions for each latent class variable based on the estimated posterior probabilities for the MIRT and MMIRT models. Student-level latent class 2 is the dominant group (.72) in the MIRT model. Note that the sum of the proportions reported in Table 6 is equal to 1. In the MMIRT model, the second level student level class 1 is the dominant group (.66).

The item parameter estimations of the final model are reported in Table 8, Table 9 and Table 10. The Mplus output provides separate slope and intercept or threshold parameters for within-group and between-groups for the MMIRT models. For this reason, the subscripts W (within-group) and B (between-group) are used to distinguish between the two levels. As illustrated in Table 6, slope (α) parameters are reported for each class at both levels. But thresholds were obtained only for the between-levels part. As described by Sen et al. (2020), the IRT discrimination parameters are equal to the slope parameters provided in the Mplus output. Nevertheless, item difficulty parameters can be obtained by dividing the threshold values for each item by the slope values. In the MIRT model, item difficulty parameters for latent class 2 appear to be higher than latent class 1.

Table 8.

Item Parameter Estimations of the Final Model for MIRT

	Latent Class 1		Latent Class 2	
Item	**α_1**	**β_1**	**$\alpha 2$**	**$\beta 2$**
1	1.18	-1.34	.55	1.01
2	.89	-3.49	.66	-.28
3	.13	-1.57	.51	2.19
4	.71	.21	-.09	-3.63
5	.39	-.43	.16	4.01
6	1.31	-.99	.10	5.35
7	1.13	.35	-.14	-1.76
8	.94	-.81	.71	.03
9	1.05	-.44	.74	.86
10	.86	-1.64	1.25	-.39
11	.92	-.79	.79	-.09
12	.94	-2.06	1.01	-.76
13	.84	-.11	.66	.80
14	.99	-.21	.56	1.36
15	.49	-4.31	.88	-.50

When item difficulty indices are analyzed, items 1,2,3,5,6,8,9,10,11,12,13,14 and 15 are lower for latent class 1 than latent class 2, i.e. they are easier. The remaining two items, items 4 and 7, are easier for latent class 2.

Table 9.

Mean scores and standard deviations for the latent classrooms in the MIRT model

Latent Class	Mean Score (Female/Male)	Standard Deviation
1	10.67(10.30/11.05)	2.38(2.35/2.35)
2	6.05(5.95/6.15)	2.39(2.32/2.45)

According to the MIRT model, the averages of male students in both implicit groups are higher than the averages of female students. In general averages, latent class 1 has a higher average than latent class 2 and it can be said that latent class 1 is more successful.

Table 10.

Item Parameter Estimates of the Final Model for Student Level

	Latent Class 1			Latent Class 2		
Item	**α_{1W}**	**α_{1B}**	**β_1**	**α_{2W}**	**α_{2B}**	**β_2**
1	1.41	1.82	-.38	.28	1.08	.40
2	1.60	1.83	-1.26	.01	-.69	1.23
3	1.33	1.58	-.65	.26	-.70	-.54
4	1.10	1.02	.51	.48	-.40	-5.04
5	.54	2.37	.58	.01	-.76	-.96
6	1.47	2.01	-.56	.34	.90	.60
7	.96	2.37	.37	.31	1.32	1.25
8	.85	-.05	4.54	-1.15	-1.19	.69
9	.79	1.87	-0.03	-1.01	1.11	.70
10	1.05	.18	-3.96	-3.32	-1.75	.94
11	.76	1.35	-.13	-1.11	1.56	-.04
12	.75	.66	-3.22	-4.70	4.46	-.13
13	.83	-.43	.66	-1.06	.35	.28
14	.81	.92	-.38	-.79	.16	4.14
15	.94	1.30	-1.50	-1.45	-.05	3.21

When item difficulty indexes are analyzed, items 1,2, 6, 7, 9, 10, 12, 14, and 15 are lower for latent class 1 than latent class 2, i.e. they are easier. The remaining six items, items 3, 4, 5, 8, 11, and 13 are easier for latent class 2. Table 11 presents the item parameter estimates of the final model for the country level of the MMIRT model.

Table 11.

Item Parameter Estimates of the Final Model for the Country Level of the MMIRT Model

	Latent Class 1			Latent Class 2		
Item	**α_1W**	**α_1B**	**β_1**	**α_2W**	**α_2B**	**β_2**
1	1.00	-.42	-1.38	1.00	.55	-3.35
2	1.31	.17	-1.02	1.31	.15	-2.36
3	.88	.25	3.90	.88	.84	-1.57
4	-.08	.23	3.93	-.08	.82	.48
5	.51	.51	1.21	.51	1.23	-.60
6	.07	-.11	-5.83	.07	-.80	1.35
7	-.09	.14	4.51	-.09	.35	1.78
8	1.18	-.26	-.59	1.18	.05	-3.16
9	1.32	-1.08	-.64	1.32	-1.79	.40
10	2.26	-.87	.53	2.26	.41	-4.15
11	1.50	-.37	.09	1.50	-.27	4.11
12	2.30	.06	-.80	2.30	.86	-2.13
13	.96	-.80	-.80	.96	-.71	.79
14	.70	-.95	-.93	.70	-2.00	.24
15	2.03	.70	-.55	2.03	1.74	-1.16

When item difficulty indexes are analyzed, items 6, 9, 11, 12, 13, and 14 are lower for latent class 1 than latent class 2, i.e. they are easier. The remaining nine items, items 1, 2, 3, 4, 5, 7, 8, 10, and 15 are easier for latent class 2. Table 12 shows the mean scores and standard deviations for the latent classes in the MMIRT model.

Table 12.

Mean scores and standard deviations for the latent classes in the MMIRT model

Latent Class	**Mean Scores**	**Standard Deviations**
1-1	6.85	2.89
1-2	6.82	2.92
2-1	7.43	3.19
2-2	7.67	3.31
1	6.84	2.89
2	7.47	3.21

Of the two latent classes for country level 1, latent class 1 has a higher mean than students in latent class 2. For country level 2, of the two latent classes, latent class 2 has a higher mean than students in latent class 1. Table 13 presents the countries included in the country-level latent classes, which is the second level in the MMIRT model.

Table 13

Countries included in the country-level latent classes of the MMIRT model

Country Level Latent Classes	
Latent Class 1	**Latent Class 2**
France	Hungary
Georgia	Türkiye
UAE	Italy
Norway	Portugal
Malaysia	Russian Federation
Finland	Israel
England	Lithuania
	Qatar
	Sweden
	Chile
	United States
	Chinese Taipei
	Hong Kong
	Korea Rep. of
	Singapore

Of the 7 countries in country level latent class 1, five are located in Europe and two in Asia. Of the 15 countries in latent class 2, eight are located in Europe, five in Asia and two in the Americas. Table 14 shows the mean scores and standard deviations for the manifest group model by gender.

Table 14.

Manifest group model mean scores and standard deviations by gender

Gender	Mean Scores	Standard Deviations
Female	7.21	3.05
Male	7.46	3.25

When the averages for the gender variable are analyzed, the averages of male students are higher than those of female students. In the MH method according to the manifest groups, analysis was made in the context of gender variable and the results obtained are shown in Table 15.

Table 15.

MH test results according to gender variable

Item	Chi Square	Alpha MH	Delta MH	Effect Size
1	4.62*	1.12	-.26	A
2	12.54***	.82	.46	A
3	63.20***	.65	1.01	B
4	8.04**	.84	.40	A
5	3.26	1.10	-.22	A
6	.02	1.08	-.02	A
7	132	1.07	-.16	A
8	.03	1.01	-.03	A
9	7.99**	.86	.35	A
10	.44	1.04	-.09	A
11	16.15***	1.22	-.48	A
12	12.27***	.82	.46	A
13	34.45***	1.35	-.70	A
14	22.71***	1.29	-.60	A
15	3.09	.91	.23	A

Annotation: ***:0.001, **:0.01, *:0.05 significance level. A: Negligible effect, B: Moderate effect, C: Large effect, indicates effect level magnitudes.

According to Table 15, it can be said that only item 3 shows DIF as a result of the MH test conducted according to the gender variable. If ΔMH>0, DIF is interpreted as DIF in favor of the focus group, ΔMH<0 as DIF in favor of the reference group, and ΔMH≅0 as no DIF (Holland & Thayer, 1986). Item 3, which had a moderate DIF effect size, displayed DIF in favor of the focal group of females. Other items show negligible level of DIF. Item 3 is a knowledge level item about finding another algebraic expression that is equivalent to an algebraic expression in algebra in mathematics. The results of the DIF analysis of the latent classes created based on item difficulties for the L0-G2 model in the MIRT model are shown in Table 16.

Table 16.

Results of the MH test for the two latent classes in the MIRT model

Item	Chi Square	Alpha MH	Delta MH	Effect Size
1	121.06***	.42	2.03	C
2	344.65***	.06	6.78	C
3	1341.28***	.03	8.51	C
4	211.66***	.30	2.80	C
5	8.17**	1.23	-.49	A
6	98.12***	.49	1.68	C
7	27.03***	.63	1.07	B
8	245.79***	3.55	-2.98	C
9	131.14***	2.51	-2.16	C
10	230.28***	3.82	-3.15	C
11	319.63***	4.59	-3.58	C
12	83.86***	2.40	-2.06	C
13	273.79***	3.96	-3.23	C
14	107.05***	2.29	-1.95	C
15	50.40***	.48	1.72	C

Annotation: ***:0.001, **:0.01, *:0.05 significance level. A: Negligible effect, B: Moderate effect, C: Large effect, indicates effect level magnitudes.

As a result of the MH test conducted for the two latent groups in the MIRT model, it can be said that all items except item 5 displayed DIF. While item 7 displayed level B DIF, the remaining 13 items displayed level C DIF. Item 5 is an algebra question at the knowledge domain and is about defining a curve with a positive slope belonging to the subject of algebra in mathematics.

The DIF analysis was evaluated at both student and country level using the MH tests. The student-level results of the MH method for the DIF analysis conducted through the item difficulty parameters in the latent groups obtained according to the L2-G2 model in the MMIRT model are given in Table 17 and the country-level results are given in Table 18.

Table 17

Results of the DIF analysis for the student-level MMIRT model

Item	Chi Square	Alpha MH	Delta MH	Effect Size
1	4.51*	.73	.74	A
2	40.60***	.38	2.28	C
3	63.46***	.31	2.73	C
4	26.25***	.35	2.49	C
5	22.17***	2.08	-1.72	C
6	44.85***	.38	2.27	C
7	.63	.87	.33	A
8	1.93	1.29	-.60	A
9	4.31*	.69	.88	A
10	0.42	.88	.29	A
11	26.69***	2.52	-2.17	C

Table 17 (continued)

Item	Chi Square	Alpha MH	Delta MH	Effect Size
12	96.14***	.16	4.34	C
13	15.94***	1.94	-1.56	C
14	15.59***	.50	1.61	C
15	.79	.82	.46	A

Annotation: ***:0.001, **:0.01, *:0.05 significance level. A: Negligible effect, B: Moderate effect, C: Large effect, indicates effect level magnitudes.

As a result of the MH test conducted for the students in the two latent groups at the first level of the MMIRT model, it was concluded that items 1, 6, 7, 8, 9, and 15 displayed DIF at a negligible effect level and items 2, 3, 4, 5, 6, 11, 12, 13, and 14 displayed DIF at a large effect level.

Table 18.

Results of the DIF analysis for the country-level MMIRT model

Item	Chi Square	Alpha MH	Delta MH	Effect Size
1	267.66***	11.32	-5.70	C
2	176.89***	19.03	-6.92	C
3	220.47***	7.79	-4.82	C
4	112.22***	4.01	-3.27	C
5	58.05***	2.43	-2.08	C
6	140.92***	4.69	-3.63	C
7	59.24***	2.76	-2.38	C
8	7.50**	1.34	-.70	A
9	1.62	1.16	-.34	A
10	25.25***	.52	1.52	C
11	5.30*	.78	.59	A
12	2.46	.85	.39	A
13	7.93**	1.36	-.72	A
14	10.14**	1.49	-.94	A
15	1.87	1.20	-.44	A

Annotation: ***:0.001, **:0.01, *:0.05 significance level. A: Negligible effect, B: Moderate effect, C: Large effect, indicates effect level magnitudes.

As a consequence of the MH test conducted for the students in the two latent country groups at the second level of the MMIRT model, it was concluded that items 8, 9, 11, 12, 13, 14, and 15 displayed DIF at a negligible effect level and items 1, 2, 3, 4, 5, 6, 7, and 10 displayed DIF at a large effect level.

Discussion

In this study, the differentiation of DIF according to manifest groups, latent classes and latent groups in which multi-levelness is taken into account was examined. For modeling both student-level and country-level data, a MMIRT model is defined. The model developed in the research utilizes the properties of IRT model, an unconstrained latent class model and a multilevel model. The first level of the model enables an opportunity to determine whether there are latent classes that differ in students' strategies for response to items. The second level of information can be used to uncover possible differences between latent classrooms in countries that may be due to curricular or pedagogical differences.

The amount of DIF items detected in both the MMIRT model and the MIRT model is higher than what would be expected from a traditional DIF analysis using manifest classes. This is because the latent group approach maximizes the differences between latent groups, resulting in a larger amount of DIF items and larger differences in item difficulties between latent groups (Samuelsen, 2005). This result is also consistent with previous research based on the use of MIRT models for DIF analysis (Cho & Cohen, 2010; Cohen & Bolt, 2005; Samuelsen, 2005).

It is seen that the amount of items with DIFs and DIF effect sizes obtained as a consequence of the analysis of the MIRT model and the MMIRT model do not exactly overlap. Standard error calculation formulas may not give accurate results in analyses in which single-level models are made within the independence assumption (Kline, 2016). For this reason, a multilevel analysis should be conducted before the analyses, and if there is multilevelness, the analyses should be conducted taking into account the multilevelness. Lee et. al. (2018) concluded that for class-specific ICC conditions, a MMIRT model is recommended instead of a single-level item response model for a clustered dataset with cluster size 20 and cluster amount 50. It was found that the same 5 items (items 2, 3, 4, 5, and 6) displayed DIF in the MIRT model and the MMIRT model. In addition, when compared with the results obtained from the DIF analysis according to the manifest groups, it was seen that only item 3 displayed DIF in all three analyses. As a consequence of the analysis based on gender, it is observed that one item displays DIF, and it becomes clear that making comparisons only according to the manifest groups is not appropriate and will lead to erroneous inferences. Uyar (2015) found that when the data are suitable for the MMIRT model, the power of the MMIRT is higher than determining DIF with manifest groups, and for this reason, when the appropriate model is used, it will be easier for experts to interpret the items displaying B and C level DIF and the reasons for the items to be biased will be determined more objectively. Finch and Finch (2013) stated that even if test takers are matched in terms of the latent trait measured, the school they attend (as a second-level variable) may lead to the presence of DIF.

When the MMIRT model is analyzed at the country level, the countries in latent class 2 generally consist of Asian countries that have achieved successful results in large-scale exams and some European countries that are also successful in these exams. In latent class 1, there are two Asian countries with low achievement levels, five European countries with moderate achievement levels. Singapore, Chinese Taipei, Korea Rep., Hong Kong and Russia, which are in latent class 2, constitute the top five in the countries participating in eTIMSS in terms of mathematics achievement. UAE and Georgia, which are in latent class 1, are in two of the last five places in mathematics achievement in eTIMSS countries. Similar results apply to the science tests. The key benefit of the MMIRT model is based on the hypothesis that the resulting latent classes represent discrete subpopulations and are not statistical artifacts of non-normality that may exist only by chance in the data (Bauer & Curran, 2003).

In this study, no multidimensional analysis was conducted due to the unidimensional structure of the data. Considering that data with multiple dimensions are frequently seen in real life situations, it is recommended to use multilevel and multidimensional MIRT models according to the data structure and it is thought that the results will be enriched. Within the scope of the research, only 2 PL models were analyzed. Analyses can be performed with 3 PL models including the effect of luck success, 4 PL models including the unlucky parameter, or simpler models (1 PL and Rasch model) and the results can be compared. In the study, data from math and science tests consisting of 15 items were used. The effect of increasing or decreasing the number of items and differentiating the selected courses can be examined. It is recommended that researchers who will conduct studies in this field should first meticulously apply preliminary analyses for the data structure, identify latent groups in accordance with the data structure and conduct DIF analysis. In addition, the results of the analysis conducted with mixture models in determining the source of DIF should be preferred even though it requires a more complex analysis because it provides more information than the results of the analysis conducted according to the manifest groups. Since there is no single correct method for determining DIF, it is recommended to apply more than one method in the studies and interpret the outputs accordingly. The duration of the analyses conducted with mixture models can be quite long depending on the dimensionality and level of the data, the selected model and quantity, and the number of items in the data. For this reason, it is recommended that the number of individuals and items should not be increased too much, but should not be set too low so as not to negatively affect the model parameters. If the parameter values are well outside the usual bounds, the analysis can be repeated by increasing the starts values. Increasing the initial values increases the time considerably. In addition, increasing the initial values slightly may not provide the desired improvement in the item statistics and the values may need to be increased further.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Author Contribution: Ömer DOĞAN: conceptualization, investigation, methodology, data curation, supervision, writing - review & editing. Burcu ATAR: conceptualization, methodology, writing - original draft, formal analysis, visualization.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as this study uses data shared with the public.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67–91.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aydemir, F. (2023). *PISA 2018 matematik ve fen bilimleri alt testlerinde değişen madde fonksiyonunun Rasch Ağacı, Mantel–Haenszel ve Lojistik Regresyon yöntemleriyle incelenmesi*. Unpublished master thesis, Gazi University, Ankara.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*(3), 338–363. https://doi.org/10.1037/1082-989X.8.3.338

Bayram, Ö. (2024). *Bir tutum ölçeği üzerinden Mantel–Haenszel ve sıralı lojistik regresyon yöntemlerine göre değişen madde fonksiyonu incelenmesi*. Unpublished master thesis, Kocaeli University, Kocaeli.

Büyüköztürk, Ş. (2018). *Veri analizi el kitabı: istatistik, araştırma deseni, SPSS uygulamaları ve yorum.* Ankara: Pegem Akademi.

Cho, S. J., (2007). *A multilevel mixture irt model for dif analysis*. Unpublished Doctoral Dissertation, University of Georgia.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture irt model with an application to dif. *Journal of educational and behavioral statistics*, *35*(3), 336–370. https://doi.org/10.3102/1076998609353111

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *83*, 278–306.

Cho, S. J., Suh, Y., & Lee, W. Y. (2015). An NCME instructional module on latent dif analysis using mixture item response models. *educational measurement:* issues and practice.

Choi, Y. J., Alexeev, N. & Cohen, A. S. (2015) Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the timss 2007 mathematics test, *International Journal of Testing, 15*(3), 239-253, DOI: 10.1080/15305058.2015.1007241

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

De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). *Differential item functioning: A mixture distribution conceptualization. International Journal of Testing, 2,* 243–276.

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement, 35, 583–603.*

Dras, L. (2023). Multilevel mixture irt modeling for the analysis of differential item functioning. Unpublished Doctoral dissertation, Brigham Young University.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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.

Finch, W. H., & Hernández Finch, M. E. (2013). Investigation of Specific Learning Disability and Testing Accommodations Based Differential Item Functioning Using a Multilevel Multidimensional Mixture Item Response Theory Model. *Educational and Psychological Measurement*, *73*(6), 973–993. https://doi.org/10.1177/0013164413494776

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement, 47*, 299-317.

Gurkan, G. (2021). *From OLS to multilevel multidimensional mixture IRT: A model refinement approach to investigating patterns of relationships in PISA 2012 data*. Unpublished Doctoral Dissertation, Boston, United States of America.

Holland, P.W. & Thayer, D.T. (1986). *Differential item performance and the Mantel-Haenszel procedure* (Technical Report No. 86–69). Princeton, NJ: Educational Testing Service.

Holland, P.W. & Thayer, D.T. (1988) Differential item performance and the Mantel-Haenszel procedure, in Wainer, H. and Braun, H.I. (Eds.): *Test Validity*, 129–145, Erlbaum, Hillsdale, NJ.

Jiao, H., & Chen, Y.-F. (2014). Differential item and testlet functioning. In A. Kunnan (Ed.), The Companion to Language Assessments (pp.1282-1300). John Wiley & Sons, Inc.

Jiao, H., Kamata, A., Wang, S. & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*, 82-100.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

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, 307-327.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kristanjansonn, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A Comparison of four methods for detecting differential item functioning in ordered response model. *Educational and Psychological Measurement, 65*(6), 935-953.

Lee, W. Y., Cho, S. J., & Sterba, S. K. (2018). Ignoring a multilevel structure in mixture item response models: impact on parameter recovery and model selection. *Applied psychological measurement*, *42*(2), 136–154. https://doi.org/10.1177/0146621617711999.

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

Lord, F.M. (1980) *Applications of item response theory to practical testing problems*, Erlbaum, Hillsdale, NJ.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2015). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-862. doi:10.3758/BRM.42.3.847.

Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute*, *22*(4), 719–748.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist, 50*(9), 741.

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

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

Paek, I., & Cho, S.-J. (2015). A note on parameter estimate comparability: Across latent classes in mixture IRT modeling. *Applied Psychological Measurement, 39*(2), 135–143. https://doi.org/10.1177/0146621614549651

Raju, N.S. (1988). The area between two item characteristic curves, *Psychometrika*, *53*(4), 495–502.

Raju, N.S. (1990) Determining the significance of estimated signed and unsigned areas between two item response functions, *Applied Psychological Measurement*, *14*(2), 197–207.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications. Revelle, W. (2023). *Psych: Procedures for psychological, psychometric, and personality research. (Version 2.3.3).* https://cran.r-project.org/web/packages/psych/psych.pdf

Revelle, W. (2023). *Psych: Procedures for psychological, psychometric, and personality research. (Version 2.3.3).* https://cran.r-project.org/web/packages/psych/psych.pdf

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

Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355–371.

Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective*. Unpublished doctoral dissertation, University of Maryland, College Park.

Sırgancı, G. (2019). *Karma rasch model ile değişen madde fonksiyonunun belirlenmesinde kovaryant (ortak) değişkenin etkisi*. Unpublished doctoral dissertation, Ankara University, Ankara.

Sen, S. (2022). *Mplus ile yapısal eşitlik modellemesi uygulamaları*. Ankara: Nobel.

Sen, S., Cohen, A., & Kim, S.-H. (2020). A short note on obtaining item parameter estimates of IRT models with Bayesian estimation in Mplus. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *11*(3), 266-282. doi: 10.21031/epod.693719

Sen, S., & Toker, T. (2021). An application of multilevel mixture item response theory model. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(3), 226-238. doi: 10.21031/epod.893149

Shealy, R. and Stout, W. (1993) A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF, *Psychometrika*, *58*(2), 159–194.

Thissen, D., Steinberg, L. & Wainer, H. (1988) 'Use of item response theory in the study of group differences in trace lines', in Wainer, H. and Braun, H.I. (Eds.): *Test Validity*, 147–169, Erlbaum, Hillsdale, NJ.

Toker, T. & Green, K. (2021). A comparison of latent class analysis and the mixture rasch model using 8th grade mathematics data in the fourth international mathematics and science study (timss-2011), *International Journal of Assessment Tools in Education 8*(4), 959–974

Unal, F. (2023). *Farklı oranlardaki kayıp verilere farklı atama yöntemleriyle veri atamanın madde tepki kuramına dayalı yöntemlerle değişen madde fonksiyonuna etkisinin incelenmesi*. Unpublished master thesis, Akdeniz University, Antalya.

Uyar, Ş. (2015). Gözlenen gruplara ve örtük sınıflara göre belirlenen değişen madde fonksiyonunun karşılaştırılması. Unpublished doctoral dissertation, Hacettepe University, Ankara.

Yalcin, S. (2018). Determining differential item functioning with the mixture item response theory. *Eurasian Journal of Educational Research 74*, 187-206

Zhang, Y. (2017). *Detection of latent differential item functioing (dif) using mixture 2pl irt model with covariate*. Unpublished doctoral dissertation. University of Pittsburgh. Pittsburgh