International Journal of Assessment Tools in Education
2025, Vol. 12, No. 3, 499–522
https://doi.org/10.21449/ijate.1531465
journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

Effects of dimensionality and covariate on items with DIF in mixture models

Ömer Doğan[1*]

[1]Ministry of National Education, Uşak, Türkiye

ARTICLE HISTORY

Received: Aug. 10, 2024

Accepted: Jan. 31, 2025

Keywords:
Mixture model,
Multidimensionality,
Covariate variable,
Differential item functioning,
Source of DIF.

Abstract: The aim of this study is to determine the differential item functioning (DIF) with a mixture model when the data set is multidimensional. The differences in determining the number of items with DIF and the source of DIF according to the status of considering dimensionality and adding the covariate to the analysis were examined. In this context, a total of 28 items of mathematics and science answered by 7965 individuals in the 3rd booklet of the electronic Trends in International Mathematics and Science Study (eTIMSS) 2019 were found to have a multidimensional structure, and the variable with the highest correlation with the data structure was determined and included in the model as a covariate. In order to select the most appropriate models for the data set, models with different numbers of latent classes belonging to the mixture model and multidimensional mixture model including the covariate were compared. Descriptive statistics of the latent classes created with the selected models were created, item parameters were examined and DIF analysis were conducted. In the light of the findings, it was determined that the number of items with DIF decreased as the model became more complex. In the model with the best knowledge criterion index, it was found that the items with DIF at the knowing level generally differed in favor of the focal group, while the items with DIF at the application and reasoning levels differed in favor of the reference group.

1. INTRODUCTION

In measurement and evaluation, various measurement tools are used according to the structure of the quality to be measured in order to measure the desired quality of people. The correct selection of these tools reduces the error rate involved in measurement and makes the evaluation more accurate. However, the skill to be measured, the number of individuals to be subjected to measurement and the characteristics of the individuals are also taken into consideration in the selection of the measurement tool. Mislevy (1993), points out that "it is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of 20th century statistics to 19th century psychology." Although various criticisms have been made similar to Mislevy's (1993) criticism, testing is the most widely used measurement tool for a reason. A test is defined as a tool or method that is obtained from a sample of examined behaviors in a particular field. It is then scored and evaluated using a standardized process. The scores obtained by administering the test (together with information from other sources) are used in the assessment. Validity is the degree to which interpretations

*CONTACT: Ömer DOĞAN ✉ 64omerdogan64@gmail.com Ministry of National Education, Uşak, Türkiye

e-ISSN: 2148-7456

of test scores are supported by evidence and theory for the proposed use of the test. Validity is, therefore, the most important basis for test evaluation and test development. The validity process involves gathering the evidence necessary for a sound scientific basis for proposed score interpretations. Validity is the interpretation of test scores for proposed uses in assessment, not the test itself. The test developer is expected to defend the validity of the proposed interpretations and uses, and it is appropriate to talk about efforts to "validate" the claims made. (American Educational Research Association [AERA] *et al*., 2014; Kane, 2006).

Validity is concerned with making inferences from test scores and involves evidence-gathering processes. One of these processes is gathering evidence about the internal structure of the test. Analyses can indicate the extent to which the relationships between test items and test components fit the structure on which the proposed test scores are based. One of the methods used to gather evidence of test internal structure is the differential item functioning (DIF). In its simplest form, DIF is defined as the difference in the probability of a correct response between two groups of test takers of the same ability level (Pine, 1977). The fact that the items of a test contain DIF is one of the important factors that reduce the validity of interpretations of test scores. Test fairness is an essential consideration when determining the validity of test scores. DIF analysis is also used to investigate whether a test is fair to subgroups of a targeted population. Features of the test itself that are unrelated to the construct being measured, or the way the test is used, can sometimes result in different meanings for scores obtained by members of different identifiable subgroups. For instance, the occurrence of DIF is purported to transpire when test takers of equivalent aptitude exhibit disparate probabilities of answering a test item correctly, as a function of group membership (AERA *et al*., 2014; Fukuhara & Kamata, 2011; Kristanjansonn *et al*., 2005; Messcik, 1995).

Regardless of the groups studied, DIF is considered a serious threat to test validity because it means that one group has an unfair advantage over another group on an item. Given the relevance of DIF to test fairness, measurement researchers have developed numerous methods to investigate DIF (De Mars & Lau, 2011). These methods have been developed from different perspectives. Gomez-Benito and Navas-Ara (2000) classified DIF detection methods based on classical test theory (CTT), item response theory (IRT), chi-square and factor analysis. In addition, Penfield and Camilli (2007) classified DIF detection methods based on overall odds ratio, LR, odds ratio differences and mean differences, and Gelin (2005) classified DIF detection methods under two main groups, namely based on manifest groups and latent variables. Manifest group refers to a community in which individuals can be differentiated according to various distinct characteristics (e.g. gender, race, language, region). Usually, methods for detecting DIF compare the functioning of items across manifest groups. However, manifest groups in which items function differently may not correspond to the true source of bias. Under a model with a latent DIF variable, DIF detection is expected to be more sensitive to this source of bias (Maij-de Meij *et al*., 2010). However, IRT-based approaches have taken a step forward over CTT-based approaches in recent years. Likewise, analyses with latent classes have also increased significantly compared to analyses with manifest groups.

The statistical foundation of item response theory (IRT) is often traced back to the seminal work of Lord, Novick, and Birnbaum (1968). Over the last 30 years, IRT measurement models and related methods such as DIF, scale linking, and computerized adaptive testing have been extensively studied and applied in achievement, ability, and skill measurement research (Reise & Revicki, 2015). Latent class analysis (LCA) and latent profile analysis (LPA) are techniques that aim to recover hidden groups from observed data. LCA and LPA are useful when you want to reduce many continuous (LPA) or categorical (LCA) variables to a few subgroups (Oberski, 2016). LCA allows the identification of specific response combinations that define latent subgroups. Thus, individuals with similar response patterns can be grouped in the same classes and subgroups with similar characteristics can be obtained. LCA aims to define homogeneous groups within a heterogeneous sample for dichotomous data, that is, to classify individuals using observed variables and to identify the observed variables or items that best distinguish the classes. For this reason, LCA is considered a very useful statistical model for analyzing quantitative data in social sciences (Dayton, 1999; Hagenaars & McCutcheon, 2002; McCutcheon, 1987). In IRT models, the latent variable is assumed to be continuous. This continuous variable is often referred to as a "latent trait" and is quantitative. In contrast, in LCA models, the latent variable is categorical and qualitative (Dai, 2009; De Ayala & Santiago, 2017; Li, 2014). Mixture IRT (MixIRT) models, which combine IRT and LCA, have been used in psychometric research to analyze item response data that may violate the basic assumptions of both modeling approaches (Rost 1990). DIF analyses using MixIRT models perform better than DIF analyses based on manifest groups and are more effective in reaching the source of DIF (Cohen & Bolt, 2005; Maij-de Meij *et al*., 2010; Samuelson, 2005). The objective of a MixIRT model is twofold. Firstly, it is required to compute item parameters and person parameters, that is to say, the latent characteristics of subjects. Secondly, it is required to estimate the latent class membership of subjects. It should be noted that in MixIRT models, a set of item parameters is estimated for each latent class. For the individual parameter, although it is assumed that a subject belongs to only one class, class membership is not known with certainty. Hence, during estimation, an individual parameter is estimated conditional on membership in each class (Dai, 2009). MixIRT models allow simultaneous calculation of the individual's ability and latent class membership and item response functions for each latent class. Individuals in each class have similar characteristics and model parameters differ across classes. Thus, IRT's assumptions of single-quality homogeneous distribution and invariance of item parameters have gained flexibility (Cho, 2013). The lowest level of MixIRT models is the Mixture Rasch model (MRM). The parameters to be estimated in this model are Rasch difficulty and class-specific ability parameters. The equation for the MRM is as follows.

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

Equation 1 relates the membership parameter *g* of a class to the item index *i*. Each individual is parameterized by an ability parameter (θ_{jg}). Where *g* is an index for the latent class, *g* = 1,....,G, *j* = 1,.....,N examinees, θ_{jg} is the latent ability of examinee j within class *g*, π_g is the proportion of examinees in each class, and b_{ig} is the difficulty parameter for item *i* in class *g*. 2-parameter logistic (2PL) MixIRT model is obtained by adding the discrimination parameter to the MRM. The 2PL MixIRT equation is as follows. In Equation 2, unlike Equation 1, a_{ig} is the discrimination parameter for item *i* in class *g*.

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

Traditional DIF applications assume that a scale is unidimensional and uses the total score to match participants from different groups on a common metric. However, in a test developed to measure multiple latent traits, if the latent traits are not highly correlated, the total score may not provide enough information to describe multidimensional distributions of latent traits. When the total score consists of two weakly correlated subscores, the relationship between the total score and one of the subscores will be severely weakened, which may reduce the representativeness of the matching variable, and thus, reduce the accuracy of the DIF assessment (Chen & Jin, 2018). An important distinction between the different MixIRT models, which is strongly related to the design of an assessment, is whether the probability of success on each item is influenced by only one of the dimensions in the model, or whether responses to an item can be modeled as depending on multiple ability dimensions at the same time. The first case is called between-item multidimensionality and the second is called within-item multidimensionality (Adams *et al*., 1997). In models with between-item multidimensionality, separate sets of items are used to measure each dimension in the model. In terms of factor analytics, these models are characterized by a simple loading structure; they can be considered as a combination of several unidimensional measurement models into a common model. The combination allows for modeling the relationships between latent ability dimensions (Hartig & Höhler, 2009). Some studies have explained both theoretically (Ackerman, 1992) and empirically (Walker & Beretvas, 2001) how item bias arises from multidimensionality and suggest that practitioners should identify all the skills that items should measure and build a complete test validity system, rather than passively accepting a unidimensional combination (Yao & Li, 2015). When data are multidimensional, standard DIF detection procedures may also be ineffective because they condition on abilities tested on a single latent trait (e.g., Mantel-Haenszel (MH) test, logistic regression, SIBTEST, likelihood ratio test). In such cases, multidimensional models are recommended as they provide both accurate parameter estimates and additional information about separate constructs (Reckase, 2009). When evaluating DIF on a multidimensional scale, all measured latent traits should be considered together. Otherwise, the measurement invariance result may be biased (Chen & Jin, 2018). This necessity is especially evident in educational tests. This is because educational and psychological tests usually consist of multiple subtests that measure multiple interrelated constructs. To broaden the scope of applications of the MixIRT model, it is therefore useful to consider a multidimensional extension of the model.

The equation for the multidimensional MixIRT model (MMixIRT) is given below. In Equation 3, in addition to Equation 2, there is a parameter θ_{jgm} which is the latent ability of examinee *j* for dimension m within class *g*.

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

One of the biggest challenges when using MixIRT models is to determine what causes the heterogeneity of the population. This is also a question that researchers often ask when using latent class modeling. This is because the latent classes identified by mixture models are difficult to interpret qualitatively. Therefore, the inclusion of potentially influential covariates in models can help to overcome the difficulties of latent class identification and improve the estimation of model parameters. In the selection of covariates, the correlation coefficients between candidate covariates and the latent variable of interest should be examined to determine the strength of the relationship (Embretson, 2007; Karadavut *et al*., 2019; Zhang, 2017). Smit *et al*. (1999, 2000) investigated the use of covariates for latent class membership in MRM and 2PL IRT models. Samuelsen (2005) further investigated the addition of covariates in the context of DIF. Dai (2013) modeled latent class membership for MRM using logistic regression with a binary covariate. Tay *et al*. (2011) conducted a real data analysis using both continuous and binary covariates as predictors of implicit class membership. In addition, Sırgancı (2019) examined the effect of covariates in MRM and DIF analysis, while Uysal Saraç (2022) examined the effect of covariates on classification and prediction accuracy. Within the scope of the research, a covariate was added to the MMixIRT model. The formula for the model where the covariate is added is as follows.

$$P\left(y_{ij}=1\middle|\theta_{jgm}\right)=\sum_{g=1}^{G}\pi_{jg/Wj}\frac{\exp[a_{ig}\left(\theta_{jgm}-b_{ig}\right)]}{1+\exp[a_{ig}\left(\theta_{jgm}-b_{ig}\right)]} \qquad (4)$$

$$\pi_{jg/Wj}=\frac{\exp(\beta_{0g}+\sum_{p=1}^{P}\beta_{pg}W_{jp}}{\sum_{g=1}^{G}\exp(\beta_{0g}+\sum_{p=1}^{P}\beta_{pg}W_{jp}} \qquad (5)$$

In Equations 4 and 5, π_{jg} is the probability that examinee *j* belongs to class *g*. Group membership *g*, has a multinomial distribution and latent groups G are modeled as functions of covariates W_{jp}

such that π_{jg} is a multinomial logit regression. β_{pg} is the class-specific effect of the covariate *p* on group membership. For identifiability, $\beta_{01} = 0$ and $\beta_{p1} = 0$ (Cho *et al*., 2013).

Studies using MixIRT models (Dai, 2013; Frick *et al*., 2015; Sırgancı, 2019; Uysal Saraç, 2022) generally use MRM, and most of these studies are simulative in terms of data structure (Li *et al*., 2009; Sırgancı, 2019; Uyar, 2015). Considering that educational tests are generally multidimensional, adding multidimensionality to MixIRT models will contribute to DIF detection. In this context, it is thought that the number of items with DIF and effect size will change when dimensionality is included in the analysis. Moreover, analyzing the data appropriately will reduce the amount of errors. Another important issue is the inclusion of covariates in the analysis. In addition to contributing to reaching the source of DIF, covariates are useful in characterizing and naming latent classes. In studies (Çepni & Kelecioğlu, 2021; Pektaş, 2018; Yalçın, 2018) where the data set is not multilevel, multidimensionality is generally not considered, and the covariate associated with this multidimensional structure is not included in the analysis. For this reason, more accurate information about the nature of DIF and items with DIF can be obtained by adding a covariate to the 2PL MMixIRT model. In scenarios where real data is employed, the data set is characterized by multidimensionality, and the analysis is designed accordingly. In addition, important information can be obtained by comparing the MixIRT model, the MMixIRT model and the MMixIRT model in which the covariate is added.

The aim of this study is to examine the effect of dimensionality and the addition of covariates when performing DIF detection with MixIRT when the data set is multidimensional. In this context, the models to be compared were examined in terms of the number of items with DIF and the source of DIF. The data set of the study consists of data from the electronic Trends in International Mathematics and Science Study (eTIMSS) 2019 3rd booklet. This is because the TIMSS tests, a large-scale international educational assessment in line with the purpose of the study, are based on multiple dimensions with three cognitive domains of mathematics and science skill sets (knowing, application, and reasoning) as well as four subject areas (numbers, geometry, algebra, and data-probability, and physics, chemistry, biology, and earth sciences). In this context, models of the mixture model, the multidimensional mixture model and the multidimensional mixture model with the addition of the covariate were created according to two, three and four latent student classes, and model fit statistics were examined. The number of items with DIF and effect sizes of the best-fitting models were compared. Furthermore, the item parameter estimates of these models were also provided. With the findings obtained, the difference caused by taking dimensionality into account and the effect of the covariates on the analysis results were investigated. In addition, the contribution of these factors in DIF studies to reaching the possible sources of DIF was examined.

2. METHOD

2.1. Research Method

The present study constitutes a descriptive research study, in which the DIF analysis results of the MixIRT models, including dimensionality and covariates, were compared. Descriptive studies describe a particular situation as completely and carefully as possible (Fraenkel *et al.*, 2012).

2.2. Sample

The research sample consisted of 8145 individuals who completed the 3rd booklet in eTIMSS 2019. However, within the scope of the study, an individual did not respond to any items. Furthermore, 179 participants were excluded from the analysis because they did not respond to the item "how often they solve problems on their own", which was selected as a covariate. The analysis was conducted using the responses of 7,965 participants.

2.3. Data

eTIMSS is an implementation of TIMSS as of 2019 and has been implemented in 22 countries excluding benchmark participants. Unlike paper-and-pencil exams, students responded to the items electronically. eTIMSS 2019 booklet 3 contains 28 four-choice multiple-choice items. The topics and cognitive domains of 12 mathematics and 16 science items are given in Table 1.

Table 1. *Subject areas and cognitive domains of the 28 multiple-choice items that constitute the data.*

Item Number	Course	Subject Area	Cognitive Domain
1	Math	Numbers	Knowing
2	Math	Numbers	Knowing
3	Math	Algebra	Knowing
4	Math	Algebra	Knowing
5	Math	Algebra	Application
6	Math	Geometry	Application
7	Math	Geometry	Reasoning
8	Math	Data and Probability	Application
9	Math	Numbers	Reasoning
10	Math	Numbers	Application
11	Math	Algebra	Knowing
12	Math	Algebra	Application
13	Science	Chemistry	Knowing
14	Science	Chemistry	Application
15	Science	Biology	Application
16	Science	Physics	Knowing
17	Science	Physics	Knowing
18	Science	Physics	Application
19	Science	Earth Science	Knowing
20	Science	Earth Science	Knowing
21	Science	Biology	Application
22	Science	Biology	Knowing
23	Science	Biology	Knowing
24	Science	Biology	Knowing
25	Science	Physics	Application
26	Science	Physics	Reasoning
27	Science	Earth Science	Reasoning
28	Science	Earth Science	Knowing

According to Table 1, five of the 12 mathematics items were in the subject area of algebra, four in the subject area of numbers, two in the subject area of geometry, and one in the subject area of data probability. In addition, five items were in the cognitive domain of knowing, five items in the cognitive domain of Application and two items in the cognitive domain of reasoning. As for the 16 science items, five of each were in biology and physics, four in earth sciences and two in chemistry. In addition, nine items were in the cognitive domain of knowing, five items in the cognitive domain of Application and two items in the cognitive domain of reasoning.

2.4. Data Analysis

Within the scope of the research, the analysis of the data set consisting of 28 items was conducted with open-source R (R Core Team, 2024) programming language and Mplus (Muthen & Muthen, 2017) program. First, the dimensionality of the data set was examined. In this context, the most used methods are Kaiser K1 rule (1960), the scree plot or parallel analysis. In the parallel analysis method developed by Horn (1965), random data sets are generated parallel to the real data set. The point where the eigenvalue of the randomly generated data is greater than the eigenvalue of the real data gives the number of factors. Within the context of the study, firstly, parallel analysis was conducted with the "paran" package (Dinno, 2018) in the R library for the items that formed the data set. After the parallel analysis, confirmatory factor analysis (CFA) was conducted for the dimensionality of the data set. After the dimensionality analysis, the ICC values of each item were obtained with the "icc" function in the "misty" (Yanagida, 2023) package in the R library to determine whether the data set exhibits a multilevel structure.

Within the scope of the research, the multidimensional data were analyzed by adding a covariate. In addition to the "About how many books are there in your home?" variable, 26 other variables such as the presence of a computer or tablet, the presence of internet, family education level, "In mathematics lessons, how often do you work on problems on your own?", "How much do you agree with these statements about learning science? I like science", "How well do you know the meaning of each of the following terms? Cut and paste" were correlated with the students' total score on the 28 items, and the variable with the highest correlation value (Pearson' r) was the variable (.33) belonging to the answers given to the item " In mathematics lessons, how often do you work on problems on your own?".

The three most commonly used information criterion indices to determine the optimal model for parameter estimation based on mixture models are Akaike's (1974) information criterion (AIC), Schwarz's (1978) Bayesian information criterion (BIC) and the sample size adjusted version of BIC (SABIC; Sclove, 1987). The equations for the criteria are given in Equations 6, 7 and 8.

$$AIC=-2\log L+2p \tag{6}$$

$$BIC=-2\log L+p*\ln(N) \tag{7}$$

$$SABIC=-2\log L+p*\ln((N+24)/2) \tag{8}$$

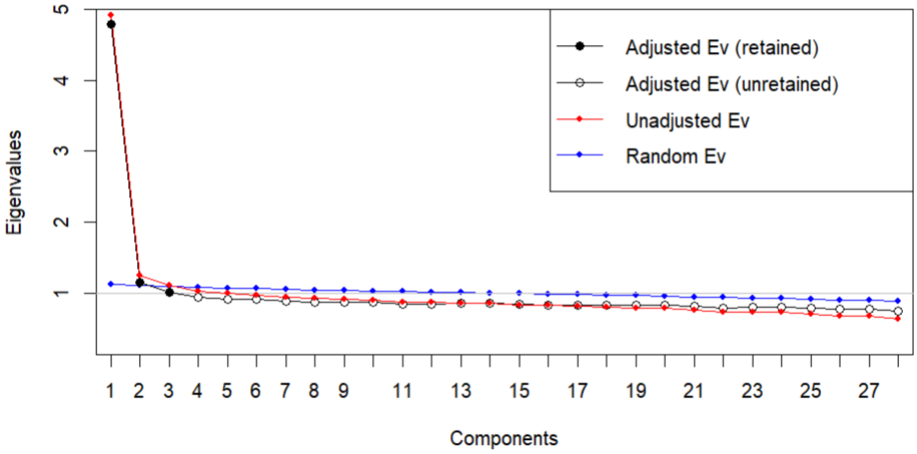
Where "L" is the likelihood function, "p" is the number of parameters to be estimated, and "N" is the sample size. AIC and BIC criteria with smaller values represent better model fit. When AIC and BIC results are different, researchers' preferences may change. McLachlan and Peel (2000) state point out the AIC criterion tends to overestimate the number of classes due to inconsistency, whereas the BIC and SABIC criteria apply more corrections to the likelihood function, and Vrieze (2012) reports that as the sample size increases, the BIC criterion tends to consistently select the correct classes. Although there are studies (Cho, 2007; Zhu, 2013) indicating that the AIC value produces more accurate results, decisions are made based on the BIC information criterion.

Mantel Haenszel (MH) method was preferred for DIF analysis. One of the reasons for this is that Diaz *et al*. (2021) argue that the MH procedure without continuity correction is the best method with respect to the Type I error rate. The other reason is that it is the most widely used method and the consequences of using it without dimensionality analysis can be seen. Therefore, many similar studies (Finch & Finch, 2013; Uyar, 2015; Yalçın, 2018) use MH to examine the DIF of latent classes determined by mixture models. Since there are two latent classes in each model, the "difMH" function of the "difR" (Magis *et al*., 2010) package from the R library was used for the analysis with the MH method in this study.

3. RESULTS

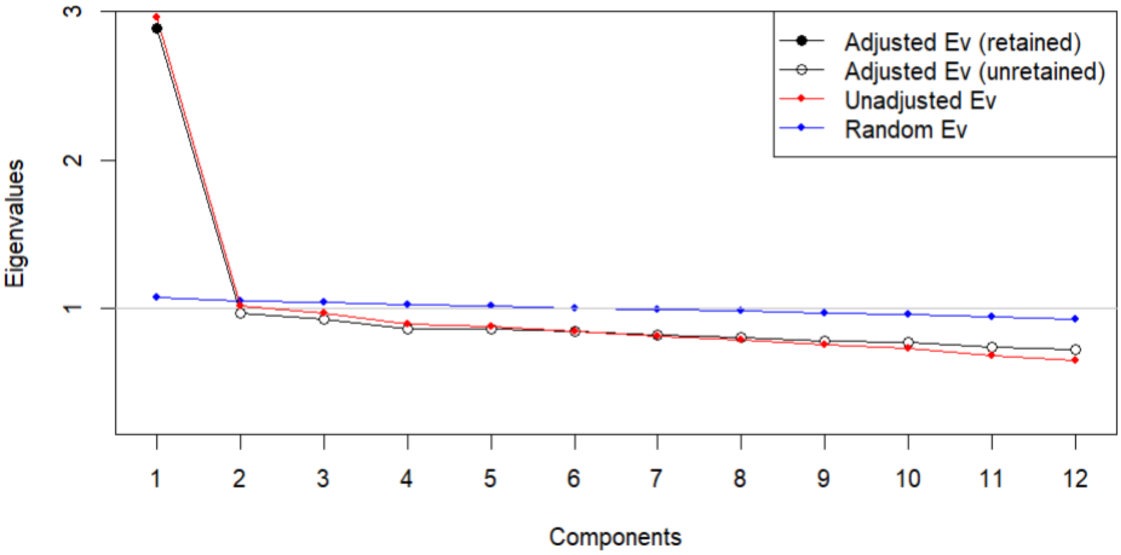
The figures and tables of the analyses described in the data analysis section are presented in the first part of this section. Within the scope of the study, the figure of the analysis performed with the “paran” package (Dinno, 2018) in the R library for the 28 items that made up the data set is shown in Figure 1.

Figure 1. *The image of the 28-item data set for the parallel analysis results.*



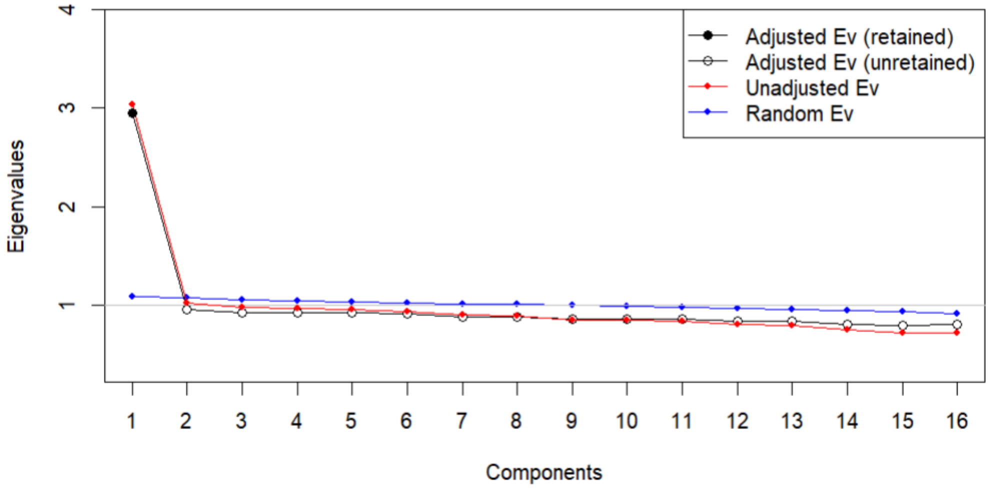
As a result of the analysis, two of the adjusted eigenvalues (retained) and two of the unadjusted eigenvalues in Figure 1 were higher than the randomly generated eigenvalues. Accordingly, it is seen that the number of factors of the data is two. Following this result, parallel analyses were conducted separately for mathematics and science items. Figure 2 displays the image of the parallel analysis results for the 12-item mathematics course.

Figure 2. *The image of the parallel analysis results for the 12-item mathematics course.*



Our analysis revealed that one of the adjusted eigenvalues (retained) and one of the unadjusted eigenvalues in Figure 2 was higher than the randomly generated eigenvalues. Accordingly, it is seen that the factor number of the data was one, that is, it was unidimensional. Figure 3 displays the image of the parallel analysis results of the 16-item science course. As a result of the analysis, one of the adjusted eigenvalues (retained) and one of the unadjusted eigenvalues in Figure 2 was higher than the randomly generated eigenvalues. Accordingly, it is seen that the number of factors of the data was one, that is, it was unidimensional. The results of the parallel analysis, which was stated to be superior to the Kaiser and scree plot methods used to determine the number of factors (Crawford *et al*., 2010; Piccone, 2009), supported the two-dimensional structure. Accordingly, the items belonging to these courses were examined by considering them as two separate dimensions with the belief that the dimensions in the test were math and science. As a result of the analysis, it was found that the items belonging to mathematics and science courses were found to be unidimensional separately.

Figure 3. *The image of the parallel analysis results for the 16-item science course.*



Following the parallel analysis, confirmatory factor analysis (CFA) was conducted for the dimensionality of the data set. The CFA was first conducted with the assumption that the dataset was unidimensional, and then, with the assumption of two-dimensional. The fit index values for both assumptions are presented in Table 2.

Table 2. *Fit index values for the unidimensional and two-dimensional structures.*

Indices	Index Value	
	Unidimensional	Two-Dimensional
AGFI	.89	.98
GFI	.90	.99
CFI	.75	.95
TLI (NNFI)	.73	.95
NFI	.74	.93
BIC	326219.01	221199.08
RMSEA	.06	.02
RMR	.01	.00

As demonstrated in Table 2, all goodness of fit index (AGFI, GFI, CFI, TLI and NFI) values obtained for the two-dimensional data structure were higher than the unidimensional data structure assumption, while the error index (RMSEA and RMR) values were smaller. Moreover, the BIC value was smaller for the two-dimensional data structure, where a smaller value means a better fit.

After the dimensionality analyses, intra class correlation (ICC) values of each item were obtained with the "icc" function in the "misty" (Yanagida, 2023) package in the R library to determine whether the data set exhibited a multilevel structure. For 28 items, the lowest ICC values were .01 and the highest .14 and the average ICC value was found to be .06. This result shows that the country level can explain 6% of the total variance. Muthén (1997) suggests that multilevel modeling should definitely be considered when the ICC value is greater than .10. Koo and Li (2016) assert that ICC values less than .50 indicate poor reliability, values between .50 and .75 indicate moderate reliability, values between .75 and .90 indicate good reliability, and values greater than .90 indicate excellent reliability. Accordingly, it was decided that a 6% variance would not be sufficient for multilevels. Table 3 displays the descriptive statistics of the 28-item data set, including measures of central tendency and dispersion, skewness and kurtosis.

Table 3. *Descriptive statistics of the data set.*

Central Tendency and Dispersion Measures	Value
Mean	13.96
Median	14
Mode	13
Standart Deviation	5.46
Minimum	0
Maximum	28
Range	28
Skewness	.2
Kurtosis	-.7

When the mean, median and mode, which are the measures of central tendency in Table 3, are examined, it is seen that these values were very close to each other. This situation indicated a normal distribution. In addition, skewness and kurtosis values vary between -1 and +1. According to Tabachnick and Fidell (2013), these values between -1.5 and +1.5 are within the acceptable range. In this case, it can be claimed that the scores did not deviate excessively from normality. The coding of the models was based on the existing usage in Mplus. Accordingly, "C" represented the latent class at the student level, and the model "C3" indicated three latent classes of students. "D-C2" represented a multidimensional model with two latent student levels, while "Cov-D-C4" represented a multidimensional model with four latent student levels where the covariate was added to the model. Table 4 displays the information criterion index values of the 9 selected models.

Table 4. *Information criterion indices for models.*

Model	LL	np	AIC	BIC	SABIC
C2	-135400.99	113	271027.98	271817.04	271457.95
C3	-135168.03	170	270676.07	271863.15	271322.92
C4	-135099.93	227	270653.86	272238.96	271517.59
D-C2	-135356.67	116	270945.33	271755.34	271386.71
D-C3	-135106.45	173	270558.89	271766.92	271217.16
D-C4	-135027.28	230	270487.07	272103.45	271406.48
Cov-D-C2	-134970.11	117	270174.20	270991.19	270619.39
Cov-D-C3	-134712.74	175	269775.48	270997.47	270441.36
Cov-D-C4	-134654.89	233	269651.12	271321.26	270659.78

LL: 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 4, the C2 model had the best information criterion index value in the MixIRT model, the D-C2 model had the best information criterion index value in the MMixIRT model, and the Cov-D-C2 model had the best information criterion index value in the MMixIRT model where the covariate was added. In the next stage, analyses were conducted using these three models. Tables 5, 6 and 7 display the item parameter values of the selected models before the DIF analyses. Table 5 displays the item parameter values of the C2 model. As demonstrated in Table 5, the discrimination index values of latent classes were generally higher for latent class 1 (LC1) than LC2. When the item difficulty index values are analyzed, it is seen that all items except items 3 and 7 were lower in LC1, that is, they were perceived more easily by students in LC1. Table 6 displays the item parameter values of the D-C2 model. As seen in Table 6, the discrimination index values of latent classes were generally higher for latent class 1 (LC1) than LC2, but the differences were quite low. When the item difficulty index values are analyzed, it is seen that all items except items 3, 7, 13, 21 and 22- were lower in LC1, that is, they were perceived more easily by students in LC1.

Table 5. *Item difficulty and discrimination index values for the C2 model.*

Item	LC1	LC1	LC2	LC2
	α1	β1	α2	β2
1	.37	-3.77	.36	.99
2	.78	-3.35	.90	.13
3	.55	.04	-.03	-3.27
4	.42	-.37	.22	3.06
5	.78	-4.47	1.12	-.31
6	.49	-2.74	.18	2.47
7	.57	.29	-.15	-1.87
8	.57	-1.61	.55	.65
9	.97	-1.03	.20	4.16
10	1.46	-1.47	.94	.13
11	1.81	-2.42	.68	.14
12	.69	-.88	.64	1.72
13	1.29	-1.96	.72	-1.48
14	1.13	-1.25	.90	.18
15	.64	.10	.31	1.77
16	.88	-.23	.50	.82
17	.67	-.66	.47	1.76
18	1.01	-2.28	.75	-.80
19	.93	-.76	.53	-.17
20	.40	-.54	.38	.25
21	1.25	-1.18	1.34	-.38
22	1.31	-1.14	.92	-.11
23	1.10	-1.16	.83	.12
24	.40	.78	.42	1.99
25	1.22	.23	.74	1.68
26	.94	-.20	.61	.90
27	1.49	-.91	1.16	-.08
28	1.23	-.72	.98	-.43

Table 7 displays the item parameter values of the D-C2 model. As seen in Table 7, the discrimination index values of latent classes were generally higher for latent class 1 (LC1) than LC2, but the differences were quite low. When the item difficulty index values are analyzed, it is seen that unlike the other models, all items were lower in LC1, that is, they were perceived more easily by students in LC1. Table 8 displays the mean scores and standard deviation values for the latent classes of the 3 best-fitting models. These scores were determined by the rule that students received 1 point for each item they responded correctly and 0 point for each item they responded incorrectly, and the highest score was 28. When the means and standard deviations of the latent classes of the models are analyzed, the LC1 latent class of the C2 model had the highest mean, while the LC2 latent class of the same model had a higher mean than the latent classes with low mean scores in the other models.

Table 6. *Item difficulty and discrimination index values for the D-C2 model.*

Item	LC1 α1	LC1 β1	LC2 α2	LC2 β2
1	.28	-.26	.24	2.41
2	1.18	-1.91	.77	.42
3	.62	-.90	-.25	-2.75
4	.38	.70	.11	3.81
5	1.54	-.74	1.11	-.17
6	.98	-1.04	.17	2.65
7	.85	-.47	-.19	-3.51
8	.87	-2.06	.55	.69
9	.74	-.63	.04	3.74
10	5.3	-2.41	1.57	-.04
11	2.04	-1.93	.41	1.02
12	.65	.14	.64	2.02
13	1.05	-.12	.56	-1.48
14	.73	-.17	.81	.55
15	.56	-1.17	.23	2.73
16	.97	-.62	.44	1.01
17	.83	-1.76	.47	1.79
18	.16	-1.94	.61	-0.41
19	1.02	-1.49	.49	-.12
20	.19	-1.52	.35	.65
21	.78	.72	1.31	-.16
22	1.00	.35	.81	.15
23	.80	-.06	.79	.40
24	.38	-.9	.31	3.09
25	1.22	-.92	.70	1.89
26	1.00	-.26	.57	1.02
27	1.37	-1.91	1.17	.04
28	1.02	-.92	.94	-.30

When the high and low achieving latent classes in the three groups were compared, the difference in mean scores between the latent classes in the D-C2 model was smaller than the others. The latent group with the lowest mean score was the LC2 latent class of the Cov-D-C2 model. Cov-D-C2, the model with the highest difference in achievement between latent groups and the lowest standard deviations, provided the best separation in terms of scores.

Table 7. *Item difficulty and discrimination index values for the Cov-D-C2 model.*

Item	LC1	LC1	LC2	LC2
	α1	β1	α2	β2
1	.93	-.83	.53	.68
2	1.12	-1.87	.65	.71
3	1.39	.68	.28	3.63
4	.45	.07	.15	3.21
5	.98	-2.83	.92	.09
6	.88	-.76	.25	1.74
7	1.37	.81	.22	4.01
8	.30	-3.20	.09	4.45
9	1.48	-.17	.40	1.98
10	.25	-3.14	.11	3.19
11	2.07	-1.19	.70	.37
12	.78	-.37	.58	2.32
13	.88	-2.78	.46	-1.53
14	.61	-2.30	.57	1.08
15	.42	.02	.10	3.26
16	.71	-.36	.32	1.91
17	.54	-.44	.31	3.30
18	1.45	-1.32	1.15	-.52
19	.60	-1.45	.25	.77
20	.27	-1.01	.41	.56
21	.45	-3.98	.91	.16
22	.95	-1.58	.68	.40
23	.63	-1.97	.61	.81
24	.39	.79	.27	3.71
25	.96	.23	.41	3.75
26	.53	-.62	.30	2.86
27	.60	-2.57	.74	.64
28	.51	-2.47	.63	.13

Table 8. *Mean scores and standard deviations of latent classes for the 3 best-fitting models.*

Model	Latent Class	Mean	Standard Deviation
C2	LC1	19.88	3.99
	LC2	11.90	4.27
D-C2	LC1	18.48	4.41
	LC2	11.13	3.93
Cov-D-C2	LC1	18.51	3.90
	LC2	10.38	3.48

After the latent classes were identified through Mplus, DIF analyses were conducted. Table 9 displays the DIF analysis statistical values and effect size for 28 items of the C2 model. If the Chi Square value was above 3.84, it was accepted that the items had DIF at .05 significance level. In addition, level A effect size was coded if the Delta MH value was in the range of 0-1, level B if it was in the range of 1-1.5, and level C if it was greater than 1.5. As a result of the DIF analysis, all items except item 4 were found to have DIF. Among the items with DIF, items

8, 10, 14, 17, 18 and 23 had A level effect size, while items 12, 13, 22 and 24 had B level effect size and the other items had C level effect size.

Table 9. *DIF analysis results for model C2.*

Item	Chi Square	Alpha MH	Delta MH	Effect Size
1	215.55	3.26	-2.78	C
2	224.86	4.97	-3.77	C
3	269.95	3.60	-3.01	C
4	.07	1.02	-.05	A
5	224.82	12.43	-5.92	C
6	312.59	4.01	-3.26	C
7	365.37	5.04	-3.80	C
8	5.62	1.20	-.43	A
9	183.49	2.69	-2.33	C
10	10.70	1.33	-.68	A
11	445.64	24.49	-7.52	C
12	59.98	1.77	-1.34	B
13	27.98	.59	1.25	B
14	12.88	.75	.69	A
15	74.50	.51	1.57	C
16	132.25	.40	2.17	C
17	6.55	1.21	-.44	A
18	7.40	1.31	-.64	A
19	158.95	.36	2.41	C
20	89.71	.49	1.68	C
21	275.66	.21	3.65	C
22	52.59	.54	1.46	B
23	25.40	.66	.98	A
24	57.88	.56	1.37	B
25	114.68	.40	2.13	C
26	116.67	.42	2.03	C
27	177.45	.30	2.82	C
28	473.31	.14	4.60	C

Table 10 displays the DIF analysis statistical values and effect size for 28 items belonging to the D-C2 model. As a result of the DIF analysis, all items except items 13, 22 and 23 were found to have DIF. Among the items with DIF, items 4, 6, 14, 15, 17, 20, 21 and 24 had level A effect size, while items 5, 8 and 12 had level B effect size, and the other items had level C effect size.

Table 11 displays the DIF analysis statistical values and effect size for 28 items of the Cov-D-C2 model. As a result of the DIF analysis, all items except items 11, 12, 13, 17, 19, 22, 26 and 28 were found to have DIF. Among the items with DIF, items 4, 6, 14, 15, 17, 20, 21 and 24 had an effect size of level A, while items 1, 14, 16, 20, 21, 24 and 27 had an effect size of level B and the other items had an effect size of level C.

Table 10. *DIF analysis results for model D-C2.*

Item	Chi Square	Alpha MH	Delta MH	Effect Size
1	344.00*	3.38	-2.86	C
2	149.10*	2.44	-2.09	C
3	190.30*	3.01	-2.59	C
4	5.61*	1.17	-0.36	A
5	36.77*	1.64	-1.16	B
6	13.22*	1.26	-.55	A
7	85.08*	2.23	-1.88	C
8	43.25*	.64	1.06	B
9	222.07*	2.61	-2.26	C
10	560.37*	.12	4.92	C
11	708.39*	8.30	-4.97	C
12	55.68*	1.63	-1.15	B
13	2.09	1.13	-.28	A
14	39.66*	1.52	-.98	A
15	37.07*	.66	.97	A
16	158.64*	.41	2.11	C
17	24.91*	.71	.81	A
18	310.21*	4.79	-3.68	C
19	202.92*	.37	2.37	C
20	6.48*	.85	.38	A
21	23.96*	.70	.83	A
22	.32	.96	.09	A
23	.16	1.03	-.06	A
24	12.24*	.79	.56	A
25	125.68*	.42	2.03	C
26	172.88*	.38	2.24	C
27	169.21*	.37	2.32	C
28	258.32*	.31	2.79	C

Table 12 displays the statistics of the DIF analyses for the most appropriate models for the data set. According to Table 12, the multidimensionality and covariate added to the MixIRT model caused a decrease in the number of items with DIF. This decrease was similar for the items with C level and total B and C level effect size.

Table 11. *DIF analysis results for model Cov-D-C2.*

Item	Chi Square	Alpha MH	Delta MH	Effect Size
1	55.29*	.57	1.34	B
2	76.50*	1.96	-1.58	C
3	99.65*	.36	2.42	C
4	21.00*	.70	.83	A
5	108.18*	2.50	-2.15	C
6	18.48*	.73	.75	A
7	51.69*	.46	-1.83	C
8	142.07*	2.37	-2.03	C
9	111.51*	.42	-2.04	C
10	903.54*	13.01	-6.03	C
11	3.04	.87	.32	A
12	3.48	1.15	-.34	A
13	1.26	1.11	-.24	A
14	62.26*	1.79	-1.37	B
15	6.71*	.82	.46	A
16	53.81*	.57	1.32	B
17	.37	1.05	-.11	A
18	255.14*	.25	3.28	C
19	4.01*	.86	.35	A
20	60.81*	.56	1.35	B
21	48.28*	1.77	-1.34	B
22	1.71	.90	.24	A
23	9.17*	1.25	-.53	A
24	37.10*	.62	1.13	B
25	5.19*	.82	.45	A
26	.00	1.00	-.01	A
27	38.44*	1.60	-1.11	B
28	.46	1.05	-.12	A

Table 12. *DIF analysis statistics using the most appropriate models for the data set.*

Model	Number of Items with DIF	Items with DIF and Effect Size Levels			Number of Items with B and C Level DIF
		A	B	C	
C2	27	6	4	17	21
D-C2	25	8	3	14	17
Cov-D-C2	20	6	7	8	15

Table 13 displays the number of latent classes for the three models, the averages of the math and science tests, the correlation value of these courses, and the overall average of the individuals in the latent classes. As shown in Table 13, there was a differentiation in the distribution of students to latent classes in the models. While the LC1 class of the C2 model had the highest overall mean, the LC2 class of the Cov-D-C2 model had the lowest mean.

Similarly, the highest mean class in mathematics was the LC1 class of the C2 model, while the lowest was the LC2 class of the Cov-D-C2 model. The mean for science was highest in the LC1 class of the Cov-D-C2 model and lowest in the LC2 class of the Cov-D-C2 model. The highest correlation for mathematics and science tests was between the latent classes of the C2 model.

Table 13. *Number of individuals, means and correlation values of latent classes in selected models and mean of covariates.*

Model	Latent Class	Number of Individuals	Mathematics Mean	Science Mean	Correlation Values	Overall Mean
C2	LC1	2050	9.18	10.70	.48	19.88
	LC2	5915	4.38	7.52	.49	11.90
D-C2	LC1	3063	8.05	10.43	.47	18.48
	LC2	4902	4.10	7.03	.37	11.13
Cov-D-C2	LC1	3504	7.74	10.77	.33	18.51
	LC2	4461	3.95	6.43	.37	10.38

Table 14 displays the DIF items of the 3 models with a common effect size of B or C, and the related subject area, subject domain and cognitive domain of these items. In addition, focal and reference group differentiations were also examined. In this context, negative values of MH statistics indicated that the item worked in favor of the reference group, while positive values indicated that the item worked in favor of the focal group (Çepni & Kelecioğlu, 2021).

Table 14. *Courses, subject areas and cognitive domains for level B and C DIF items.*

Items with DIF with a common effect size of level B or C						
	Focal Group					
Item	C2	D-C2	Cov-D-C2	Course	Subject Area	Cognitive Domain
1	-	-	+	Math	Numbers	Knowing
2	-	-	-	Math	Numbers	Knowing
3	-	-	+	Math	Algebra	Knowing
5	-	-	-	Math	Algebra	Application
7	-	-	-	Math	Geometry	Reasoning
9	-	-	-	Math	Numbers	Reasoning
16	+	+	+	Science	Physics	Knowing
27	-	-	-	Science	Earth Science	Reasoning

As demonstrated in Table 14, when the advantage of the common items of six mathematics and two science courses with B or C level DIFs to the focal and reference groups is examined, it is seen that all of the common items in the mathematics sections of the C2 and D-C2 models provided advantage to the reference group and one of the common items in science provided advantage to the focal group. According to the Cov-D-C2 model, four of the six mathematics items were advantageous for the reference group, while one of the two items in science was advantageous for the focal group and the other for the reference group. When the subject areas were examined, there was a contrast in the differentiation of the focal and reference group between the models in algebra and numbers in mathematics. When the cognitive domains were analyzed, the difference in the level of knowing stood out.

4. DISCUSSION and CONCLUSION

In the research, using the data in the third booklet of eTIMSS 2019, latent classes were created with MixIRT, MMixIRT and MMixIRT models with the addition of a covariate, DIF analysis was performed and the variables that could be the source of DIF were examined. The models with the best information criterion value of these models were selected and all selected models were the ones with two latent classes. Sen and Cohen (2019) found that the number of latent classes selected in the MixIRT studies they analyzed varied between one and ten, and most of the studies used models with two latent classes (63%). Although the most common model in MIRT studies is the Rasch mixture model (50%) (Sen & Cohen, 2019), the 2PL model was used in this study to examine item discrimination index values. When the item discrimination index values are examined, it is seen that the three models that provided the best fit were close to each other. In addition, considering the discrimination index values, it was observed that the items in LC2 latent classes were lower and some of them had negative discrimination values compared to LC1 latent classes. Choi *et al*. (2015) found that the group with lower achievement level had higher discrimination values in most of the items in their study. However, it was observed that the discrimination values of the group with higher achievement level were higher in the multiple-choice items in the composite test used.

In the latent classes created for the models, the number of students in the latent classes with low achievement levels was higher in all three models. In addition, these latent classes had significantly lower mean scores in both courses. In Finch and Finch's (2013) study, two of the three latent classes were found to have high mean achievement and low mean achievement in the selected courses, while the third latent class was found to have the highest mean achievement in one course and the lowest mean achievement in the other course. This reveals potential omissions of comprehensive analyses and suggests that when multiple constructs are assessed in a single test session, MMixIRT can provide more information about uniform DIF by examining multiple dimensions simultaneously (Finch & Finch, 2013). The inclusion of multiple dimensions provides a more complete characterization of test takers by simultaneously using their relative proficiency on multiple constructs to reflect the real-world contexts in which students learn and are assessed. In this study, the two-dimensional structure across items was considered and the inclusion of dimensionality improved the knowledge criterion index. Many studies (Bulut & Suh, 2017; Choi & Wilson, 2015; Hartig & Höhler, 2009; Yao & Li, 2015) have addressed the concept of dimensionality in the context of multidimensional IRT and the results of the analysis emphasized the importance of considering dimensionality. In the context of dimensionality, Gürdil (2023) found that the results of multidimensional estimation obtained lower Type I error and higher statistical power ratios than one-dimensional estimation in all DIF detection methods. Finch and Finch (2013) reported that the results of their analysis showed that the multidimensional model provided more complete information about the nature of the DIF than separate one-dimensional models.

As a covariate in the study, the responses to the item on ’How often do they solve math problems on their own?’ were used. This variable was not related to socioeconomic status, which was often used, but was an individual preference. Although problem solving on its own was perceived as finding a correct result, it was an action that encompassed a broader mental process and skills, and the problem-solving process involved conducting research through controlled activities to reach a goal that was clearly designed but not immediately attainable (Altun, 2004; Polya, 1957). For this reason, with the addition of the covariate, the advantageous group in DIF items differed in some items compared to other models. Sırgancı (2019) found that with the addition of a discrete covariate to the MRM, the power and accuracy of DIF detection was higher than the MRM. In other words, if there was prior knowledge about the variables that were thought to be the source of DIF, including these variables as covariates in the MixIRT model showed that more accurate results could be obtained in DIF identification. Smit *et al*. (1999; 2000) suggested that latent class assignment could benefit significantly from the inclusion of dichotomous covariates that were moderately or strongly related to the latent class variable. In line with these results, Li *et al*. (2016) found in their study that correctly specifying both dichotomous and continuous covariates in MRM led to a moderate increase in the correct classification rate. Similarly, Choi *et al*. (2015) reported that adding a covariate had the potential to improve the interpretation of differences between groups. In the light of this information, it is estimated that the Cov-D-C2 model, to which the covariate is added, classifies latent classes more accurately. This suggests that the results of the DIF analysis conducted with the Cov-D-C2 model are more reliable. The multidimensionality and covariate added to the MixIRT model have caused a decrease in the number of items with DIF. Although the current decrease does not suggest that making additions to the model leads to more accurate identification of items with DIF, the fact that it has better information criterion criteria may allow inferences to be made. Nevertheless, this is an issue that needs to be examined in tests for which items are available. On the other hand, as Gierl *et al*. (2000) suggested two decades ago, the priority should be to explain the causes of DIF due to the trend towards using more theory-oriented tests for the existence of DIF rather than examining statistically marked DIF items in an effort to find an explanation for DIF that was large enough to be statistically significant. Huang (2010) stated that the results of her research showed that DIF was not a fixed characteristic of any test item, and that DIF was a function of the use and interpretation of test scores, not an inherent feature of the test. Yıldıztekin (2020) found that taking dimensionality into account resulted in fewer items with DIF than the analyses conducted without taking dimensionality into account. Huang (2010) examined DIF as a result of language, curriculum and culture, and found that when dimensionality was included, the number of items with DIF decreased for all comparisons. Liaw (2015) also reached a similar conclusion. In Cho and Cohen's (2010) study, adding a covariate did not change the number of items with DIF.

The items with DIF and their numbers in the findings of the studies differ according to the models. Expert opinion, which was one of the practices to determine whether these items were indeed DIFs, was not conducted because the items could not be accessed. Therefore, the inferences made below were based on the findings. When the DIF results of the Cov-D-C2 model, which provided the best fit, were examined, it was seen that the majority of the items with DIF in favor of the focal group were at the knowing level, while the DIF worked in favor of the reference group in the differences at the application and reasoning levels. According to the results obtained, three of the eight mathematics items identified as having DIF were at the knowing and application levels and two were at the reasoning level. While two of the three items at the knowing level were in favor of the focal group, and one was in favor of the reference group, all five items at the application and reasoning levels were in favor of the reference group. In addition, three of the seven science items identified as DIFs were at the knowing and practicing levels and one was at the reasoning level. While all the three items at the knowing level were in favor of the focal group, three of the four items at the application and reasoning level were in favor of the reference group and one of the four items at the application and reasoning level were in favor of the focal group. While C2 and D-C2 models gave the same results for seven items with B and C level DIFs common to three models, Cov-D-C2 model gave results in favor of the focal group in two items at the knowledge level. Up to now, stating the results one of the reasons for this situation could be the idea that the selected covariate was effective in the formation of the focal and reference groups, and the items in which the reference group differed at the application and reasoning level, which was the upper level, and the focus group differed at the knowledge level, which was the lower level, with the effect of problem solving frequency, were revealed in this way. Yalcin (2018) conducted DIF analysis in two different ways, and in the first study, no item with DIF was found in favor of the focal group. The reason for this was that the three countries that made up the sample were successful, middle-achieving and low-achieving countries. In her analysis according to the MH method, seven of the ten items were found to have DIF in favor of the reference group and three in favor of the focal group.

One of the objectives of the study was to reach the source of DIF, and when dimensionality and covariate conditions were added to the model, it was seen that there was a difference in the number of items with DIF, the level of effect and the affected group (focal or reference). Accordingly, it was found that the manifest variable (e.g. gender, race) included in many studies

(Cohen & Bolt, 2005; Cohen *et al*., 2005; Doğan & Atar, 2024; Maij-de Meij *et al*.,2010; Samuelson, 2005) might not provide accurate information about DIF and MIRT models were more reliable. With the extension of the model, it became easier to explain the reason why the item with DIF in which course, subject and cognitive domain differed in favor of which group. Thus, it was revealed that the focal group differed at the knowledge level and the reference group differed at the other two levels. In this way, the reason for the correctly determined differentiation can be discussed by field experts (e.g. curriculum development, teaching principle methods, mathematics and science field experts, teachers, administrators, ministries of education, politicians) and renewal or improvements can be made on issues such as curriculum, teaching techniques or assessment and evaluation processes.

Within the scope of this study, eTIMMS 2019 data were used. However, since it was not possible to examine the items in the study, the items showing DIF could only be analyzed in the context of the course, subject area and cognitive domain. Researchers who will study in this field can reach more comprehensive results if they use the data whose items they can access. An expert evaluation process can be carried out with the accessed items. In addition, only 2PL model was used in the study. In addition to the frequently used Rasch model, such studies can be conducted with 3PL and 4PL models. It is important to examine the data set very carefully before the application and to make an appropriate analysis in order for the research to serve the purpose. For example, since the data set of this study was single-level, multilevel analyses were not conducted. In studies where covariates will be included, diversification of covariates may give a more detailed idea about the source of DIF. Very few studies in this field have been conducted using open-ended items. For this reason, different item types can be used in MIRT or more comprehensive analyses.

Declaration of Conflicting Interests and Ethics

The author declares 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.

Orcid

Ömer Doğan https://orcid.org/0000-0001-5169-520X

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. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Adams, R., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23. https://doi.org/10.1177/0146621697211001

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

Altun, M. (2008). *Matematik öğretimi [Math teaching]*. Aktüel.

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.

Baker, F. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation

Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education, 2*, 1-14. https://doi.org/10.3389/feduc.2017.00051

Chen, H.F., & Jin, K.Y. (2018). Applying logistic regression to detect differential item functioning in multidimensional data. *Frontiers in Psychology*, *9*, 1-11. https://doi.org/10.3389/fpsyg.2018.01302

Cho, S.J. (2007). *A multilevel mixture IRT model for DIF analysis* [Unpublished Doctoral Dissertation]. University of Georgia.

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

Choi, Y.J., 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, https://doi.org/10.1080/15305058.2015.1007241

Choi, I.H., & Wilson, M. (2015). Multidimensional classification of examinees using the mixture random weights linear logistic test model. *Educational and Psychological Measurement*, *75*(1), 78–101. https://doi.org/10.1177/0013164414522124

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

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

Crawford, A.V., Green, B.S., Levy, R., Lo, W.J., Scott, L., Svetina, D., & Thompson, M.S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurements*, *70*(6), 885-901. https://doi.org/10.1177/0013164410379332

Çepni, Z., & Kelecioğlu, H. (2021). Detecting differential item functioning using SIBTEST, MH, LR and IRT methods. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(3), 267-285. https://doi.org/10.21031/epod.988879

Dai, Y. (2009). *A mixture Rasch model with a covariate: A simulation study via Bayesian Markov Chain Monte Carlo estimation* [Doctoral dissertation, University of Maryland]. Digital Repository at the University of Maryland. http://hdl.handle.net/1903/9926

Dai, Y. (2013). A mixture Rasch model with a covariate: A simulation study via Bayesian Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *37*, 375-396. https://doi.org/10.1177/0146621612475076

Dayton, C.M. (1999). *Latent class scaling analysis. Sage university paper series on quantitative applications in the social sciences*. Sage.

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

DeMars, C.E., & Lau, A. (2011). Differential item functioning detection with latent classes: how accurately can we detect who is responding differentially? *Educational and Psychological Measurement*, *71*(4), 597-616. https://doi.org/10.1177/0013164411404221

Diaz, E., Brooks, G., & Johanson, G. (2021). Detecting differential item functioning: Item response theory methods versus the Mantel-Haenszel procedure. *International Journal of Assessment Tools in Education* *8*(2), 376–393. https://doi.org/10.21449/ijate.730141

Dinno, A. (2018). *paran: Horn's test of principal components/factors*. R package version 1.5.2, https://CRAN.R-project.org/package=paran

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

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

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

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). *How to design and evaluate research in education* (8th ed.). Mc Graw Hill.

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

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement 35*(8) 604–622. https://doi.org/10.1177/0146621611428447

Gelin, M.N. (2005). *Type I error rates of the DIF MIMIC approach using Joreskog's covariance matrix with ML and estimation* [Unpublished doctoral dissertation]. The University of British Columbia.

Gierl, M.J., Jodoin, M.G., & Ackerman, T.A. (2000). *Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Regression when the proportion of DIF items is large* [Paper presentation]. The Annual Meeting of the American Educational Research Association, New Orleans, LA.

Gomez-Benito, J., & Navas-Ara, M.J. (2000). A comparison of χ^2, RFA and IRT based procedures in the detection of DIF. *Quality and Quantity, 34*(1), 17-31. https://doi.org/10.1023/A:1004703709442

Gürdil, H. (2023). *The use of multidimensional item response theory models in controlling differential item functioning* [Unpublished doctoral dissertation]. Ankara University.

Hagenaars, J.A., & McCutcheon, A.L. (Eds.). (2002). *Applied latent class analysis*. Cambridge University Press.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*(2-3), 57-63. https://doi.org/10.1016/j.stueduc.2009.10.002

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrica, 30*(2), 179-185. https://doi.org/10.1007/BF02289447

Huang, X. (2010). *Differential item functioning: The consequence of language, curriculum, or culture?* [Unpublished doctoral dissertation]. UC Berkeley.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151. https://doi.org/10.1177/001316446002000116

Kane, M.T. (2006). Validation. In R.B. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). American Council on Education/Praeger.

Karadavut, T., Cohen, A.S., & Kim, S.H. (2019). Mixture rasch model with main and interaction effects of covariates on latent class membership. *International Journal of Assessment Tools in Education, 6*(3), 362-377. https://doi.org/10.21449/ijate.592789

Koo, T.K., & Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

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. https://doi.org/10.1177/0013164405275668

Li. T. (2014). *Different approaches to covariate inclusion in the mixture Rasch model* [Unpublished Doctoral dissertation]. University of Maryland.

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

Li, T., Jiao, H., & Macready, G.B. (2016). Different approaches to covariate inclusion in the mixture Rasch model. *Educational and Psychological Measurement*, *76*(5), 848-872. https://doi.org/10.1177/0013164415610380

Liaw, Y.L. (2015). *When can multidimensional item response theory (MIRT) models be a solution for differential item functioning (DIF)? A Monte Carlo Simulation Study* [Unpublished doctoral dissertation]. University of Washington.

Maij-de Meij, A.M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, *45*(6), 975-999. https://doi.org/10.1080/00273171.2010.533047

McCutcheon A.C. (1987). *Latent class analysis*. Sage.

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

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), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Mislevy, R.J. (1993). Foundations of a new test theory. In N. Frederiksen, R.J. Mislevy, and I.I. Bejar (Eds.), *Test theory for a new generation of tests*. Erlbaum.

Muthén, B. (1997). Latent variable modeling of longitudinal and multilevel data. *Sociological Methodology*, *27*, 453–480. http://www.jstor.org/stable/271113

Muthén, L.K., & Muthén, B.O. (2017). *Mplus: Statistical Analysis with Latent Variables: User's Guide* (Version 8). Authors.

Oberski, D.L. (2016). Mixture models: Latent profile and latent class analysis. In J. Robertson, & M. Kaptein (Eds.), *Modern statistical methods for HCI* (pp. 275-287). (Human–Computer Interaction Series). Springer. https://doi.org/10.1007/978-3-319-26633-6_12

Pektaş, S. (2018). *The effects of differential item functioning determination methods on test parameters estimates, decision studies, g and phi coefficients* [Unpublished doctoral dissertation]. Gazi University.

Penfield, R.D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (pp. 125-167). Elsevier Science & Technology. https://doi.org/10.1016/S0169-7161(06)26005-X

Piccone, A.V. (2009). *A comparison of three computational procedures for solving the number of factors problem in exploratory factor analysis* [Unpublished Doctoral Dissertation]. University of Northern Colorado. https://digscholarship.unco.edu/dissertations/228/

Pine, S.M. (1977). Applications of item characteristic curve theory to the problem of test bias. In D.J. Weiss (Ed.), *Applications of computerized adaptive testing: Proceedings of a symposium presented at the 18th Annual Convention of the Military Testing Association* (Vol. 77, No. 1, pp. 37–43). University of Minnesota, Department of Psychology, Psychometric Methods Program.

Polya, G. (1957) *How to solve it. a new aspect of mathematical method* (2nd ed.). Princeton University Press.

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reckase, M.D. (2009). *Multidimensional item response theory*. Springer.

Reise, S.P., & Revicki, D.A. (Eds.). (2015). *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge/Taylor & Francis Group.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective* [Doctoral dissertation, University of Maryland, College Park]. ProQuest Dissertations and Theses Global.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2), 461–464. http://www.jstor.org/stable/2958889

Sclove, L. (1987) Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333-343. http://dx.doi.org/10.1007/BF02294360

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

Sırgancı, G. (2019). *The effect of covariant variable on determination of differential item functioning using mixture rasch model* [Unpublished doctoral dissertation]. Ankara University.

Smit, A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed rasch models. *Methods of Psychological Research Online*, *4*(3), 1-13.

Smit, A., Kelderman, H., & van der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research Online*, *5*, 31-43.

Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Tay, L., Newman, D.A., & Vermunt, J.K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods*, *14*(1), 147-176. https://doi.org/10.1177/1094428110366037

Uyar, Ş. (2015). *Comparing differential item functioning based on manifest groups and latent classes* [Unpublished doctoral dissertation]. Hacettepe University.

Uysal Saraç, M. (2022). *Comparison of covariate including approaches to mixture rasch model in terms of classification and estimation accuracy* [Unpublished doctoral dissertation]. Gazi University.

Vrieze, S.I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228-243. https://doi.org/10.1037/a0027127

Walker, C.M., & Beretvas, S.N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: a cognitive explanation for DIF. *Journal of Educational Measurement*, *38*(2), 147–163. https://doi.org/10.1111/j.1745-3984.2001.tb01120.x

Yalçın, S. (2018). Determining differential item functioning with the mixture item response theory. *Eurasian Journal of Educational Research*, *18*(74), 187-206.

Yanagida, T. (2023). *misty: Miscellaneous functions for statistical analysis and data management* (Version 0.5.2) [Computer software]. https://doi.org/10.32614/CRAN.package.misty

Yao, L., & Li, F. (2015). A DIF detection procedure in multidimensional item response theory framework using MCMC technique. *International Journal of Quantitative Research in Education*, *2*(3), 285-304.

Yıldıztekin, B. (2020). *The investigation of test dimension effect to differential item functioning under different conditions* [Unpublished doctoral dissertation]. Hacettepe University.

Zhang, Y. (2017). *Detection of latent differential item functioning (DIF) using mixture 2PL IRT model with covariate* [Unpublished doctoral dissertation]. University of Pittsburgh.

Zhu, X. (2013). *Distinguishing continuous and discrete approaches to multilevel mixture IRT models: A model comparison perspective* [Unpublished doctoral dissertation]. University of Maryland.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Lawrence Erlbaum Associates, Inc.