ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 76-94

Comparison of Methods Used in Detection of DIF in Cognitive Diagnostic Models with Traditional Methods: Applications in TIMSS 2011*

Büşra EREN** Tuba GÜNDÜZ*** Şeref TAN****

Abstract

This study aims to compare the Wald test and likelihood ratio test (LRT) approaches with Classical Test Theory (CTT) and Item Response Theory (IRT) based differential item functioning (DIF) detection methods in the context of cognitive diagnostic models (CDMs), using the TIMSS 2011 dataset as a retrofitting study. CDMs, which have a significant potential when determining the DIF and their contribution to validity, can give confidence under the strong methodological background condition is met. Therefore, it is hoped that this study will contribute to the literature to ensure the correct usage of CDMs and evaluate the compatibility of these new approaches with traditional methods. According to the analysis results, thirty-one items showed differences between the cognitive diagnosis assessments and the traditional methods. The item with the largest DIF was found in the Raju Unsigned Area Measures technique in IRT, whereas the item with the lowest DIF was found in the Wald test technique developed for CDMs. In general, the analyses show that methods not based on CDMs detect more items with DIF, but the Wald test and LRT methods based on CDMs detect fewer items with DIF. This study conducted DIF analyses to determine the test's psychometric properties within the framework of CDMs rather than the source of the bias. Researchers can take the study one step further and make more specific assessments about the items' bias regarding the test structure, test scope, and subgroups. In addition, DIF analyses in this study were carried out using only the gender variable, and researchers can use different variables to conduct studies specific to their purpose.

Keywords: Cognitive diagnosis models, large scale assessment, differential item functioning

Introduction

Cognitive Diagnostic Models (CDMs) are psychometric models that provide detailed information about examinees' mastery of interrelated but separable attributes (Hou et al., 2014). Rather than dealing with students' positions on a continuous latent variable as Item Response Theory (IRT) does, CDMs predict a profile of categorical latent attributes. The term "attribute" is used to refer to latent variables in the study because latent variables assume that the items in the measurement tool may be related to one or more latent variables, which are referred to by various names in the literature, such as ability, and skill (Paulsen et al., 2020; Ravand & Baghaei, 2019).

CDMs, which provide examinees with finer-grained diagnostic information, enable them to be classified according to their mastery profiles (DiBello & Stout, 2007). In this classification, the correct response to the item indicates that the student has the necessary attributes, represented by "1" in the Q-matrix entries. Otherwise, these entries are "0" (Rupp et al., 2010). This matrix, which is essential in

*The present study is a part of PhD Thesis conducted under the supervision of Prof. Dr. Şeref TAN and prepared by Büşra EREN.

** Instructor, National Defence University, Department of Educational Sciences, Balikesir-Türkiye, busra_karaduman@yahoo.com, ORCID ID: 0000-0001-7565-1025
*** Res. Assist. PhD., Mugla Sitki Kocman University, Faculty of Education, Mugla-Türkiye, tuba.karacan@yahoo.com, ORCID ID: 0000-0002-0921-9290
**** Prof. Dr., Ankara-Türkiye, sereftan4@yahoo.com, ORCID ID: 0000-0002-9892-3369

To cite this article:
Eren, B., Gündüz, T., & Tan, Ş. (2023). Comparison of methods used in detection of DIF in cognitive diagnostic models with traditional methods: Applications in TIMSS 2011. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 76-94. https://doi.org/10.21031/epod.1218144

Received: 12.12.2022
Accepted: 6.03.2023

determining the profiles of students regarding the attributes that do plan to measure with the test and which is confirmatory, is a common point of CDMs. It maps the attributes required by the items in a multidimensional way by placing them in rows and the attributes in columns, using a simple or complex load structure (de la Torre & Minchen, 2014; Rupp et al., 2010). The validity of the findings gathered from the students' responses increases when the items and attributes in the matrix correctly match within the framework of the relevant structure (Ravand & Baghaei, 2019). Therefore, identifying the Q-matrix used in CDMs becomes essential in testing development when considering its accuracy and design (Kang et al., 2018). When this step omits, the studies' findings indicate biases in item parameters and problems in student classification (de la Torre, 2008; de la Torre & Chiu, 2016). Bias in the parameters is among the factors affecting the validity. It may occur when the scores of students in different subgroups contain systematic errors (Camilli & Shepard, 1994).

It is stated in the literature that many important statistical routines are needed to ensure appropriate uses and interpretations and to unlock the potentials of CDMs, such as the procedure for detecting differential item functioning (DIF), which can be used to determine item parameter bias (Ma et al., 2021; Paulsen et al., 2020). DIF has been described traditionally as "the probability of students with the same total test score or ability level but in different groups to correctly respond to an item when the variable is unrelated to the construct of interest" (Hou et al., 2014). For example, suppose an item has a systematic advantage favouring the female group. The item might be biased since the item response function differs between the female and male groups. A problem will arise regarding the validity of scores obtained from the test since different properties are mixed with the property to be measured. Researchers should identify and examine biased items to eliminate the problem and perform proper measurement procedures (Lee et al., 2021).

DIF is as essential in CDMs as it is in traditional approaches. "Traditional approaches make rankings at the latent ability level, while CDMs focus on the change in correct response probability regarding the responses given to an item by students in various groups but with similar attribute mastery profiles" according to the difference between the two types. In other words, DIF is defined according to CDMs as "an effect in which the probability of answering an item is different correctly for students with the same attribute mastery profile but from different observed groups" (Hou et al., 2014).

CDMs, a multidimensional model that has been increasingly popular in recent years, are used to obtain diagnostic information about students' strengths and weaknesses. This information, along with feedback opportunities for teachers and programs, provides students with opportunities for individualized learning support that compensates for learning deficiencies. In addition to this contribution, DIF detection, one of the most important statistical routines for ensuring proper usage and interpretation, appears to be a helpful method, mainly when dealing with the issue of validity, which is a problem with traditional methods (Akbay, 2021). Therefore, detecting DIF has become a standard procedure in psychometric analyses. The presence of items with DIF can worsen the predictions of attributes (Paulsen et al., 2020) and disturb the attribute profiles, causing problems in comparing latent classes between groups (Hou et al., 2014). DIF analysis is also necessary to examine parameter or configural invariance (Zumbo, 2007). According to attribute profiles, item responses that must independently condition are considered invariant. As a result, DIF analysis is critical for determining whether attribute-item interactions between groups are invariant (Hou et al., 2014).

There is little research on determining DIF in CDMs in the literature. Milewski and Baron (2002) applied DIF to individual skill performance and compared the results of four DIF methods without considering item biases. Zhang (2006) compared traditional methods limited to uniform DIF (MH and SIBTEST) at the level of attribute and item in determining DIF and took into account different simulation conditions using deterministic noisy “and” gate (DINA). When the findings were examined, when the conditions related to the test scores and the attribute profiles were taken into account, it was seen that the matching in the attribute profiles resulted in lower Type I error and higher power rates compared to the test scores, but both methods showed poor performance. Li (2008) extended High-Order DINA (HO-DINA), which was developed by de la Torre and Douglas (2004), to examine DIF and differential attribute functioning (DAF) simultaneously. The new approach used the MCMC algorithm, including Gibbs sampling, to

estimate and compare the model with the traditional MH technique regarding Type I error and power rates under different conditions. In addition, the presented new approach was examined in real-life conditions using a mathematics test. In their simulation study, Hou et al. (2014) developed a new technique (Wald test) to analyze uniform and non-uniform DIF in CDMs. With a simulation study, Liu et al. (2019) investigated the performance of the Wald test in determining DIF using various covariance matrices. To determine DIF in CDM, Hou et al. (2020) utilized the Wald test formulations. The performances of the items in the real dataset were investigated under various simulations, and the compatibility of the attributes' classifications was evaluated when saturated and reduced models were used. In the CTT, IRT, and CDM framework, Akbay (2021) investigated the test's psychometric properties using DIF determination methods (i.e., MH, Raju area measures, and Wald test for DIF). DIF flagging patterns of three different DIF detection methods were observed when real data from a large-scale assessment (TEOG) were retrofitted. The data was collected using Booklets A and B, and DIF analyses were conducted in subgroups based on gender and booklet-type variables. Finally, the studies of Ma et al. (2021) changed the assumptions of the multi-group G-DINA model (MG G-DINA). They developed the MG-G-DINA model for DIF detection to reveal that students in different groups could use the same or different attributes in various ways and compared the performance of this model with the likelihood ratio test (LRT) and Wald test.

Even though there are methods for determining DIFs, studies in the literature suggest that a more effective approach for estimating DIFs is still worth investigating. Because most CDM research has been constrained to research settings over the last decade, many psychometric questions about DIFs related to these models remain unanswered. These non-diagnostic assessments have been retrofitted into CDMs to give detailed information while being examined with traditional models. These are crucial steps in shifting from single-score reporting to CDMs that provide more thorough feedback. The retrofitting is thought to be useful in determining the DIF in order to provide detailed inferences about the students and to provide appropriate use and interpretation of CDMs in the context of the validity and reliability of the inferences regarding the test scores, given exam investments (Terzi & Sen, 2019). Searching for meaning in an evaluation without making assumptions about validity will not give the promised benefit or have the desired influence on educational policies. Therefore, the importance of performing CDM analyses with large-scale datasets should be emphasized in the literature because the differentiation of the exam language, the differences between cultures, or the differences in demographic variables such as gender cause some changes in students' performance. Due to these changes, it will be important to consider the situations that may affect student performance in examining scores (Asil & Gelbal, 2012, Odabas, 2016).

Considering the contributions mentioned in the literature regarding the determination of DIF and its validity, CDMs, which have significant potential, can give confidence, provided that the methodology is sound. As a result, in this study, the Wald test (Hou et al., 2014; Ma et al., 2021) and LRT (Ma et al., 2021), which are based on the MG G-DINA model used in cognitive diagnostic assessments, and Mantel-Haenszel (MH; Mantel & Haenszel, 1959) and logistic regression (LR; Swaminathan & Rogers, 1990) methods, which are based on CTT, and Lord's $\chi 2$ (Lord, 1980) and Raju's unsigned area measures Raju (1988), which are based on IRT methods were compared by using a large-scale dataset TIMSS (Trends in International Mathematics and Science Study) 2011 to ensure the correct use of CDMs. The approaches' compatibility and DIF's effect on CDM were examined using these comparisons. For this purpose, the existence of many studies showings that items with DIF in the bias analyses performed between gender groups, especially in numerical fields such as mathematics, played an important role in the selection of gender as the DIF variable within the scope of the study.

Since there is no single effective method for detecting DIF, using more than one method in the literature is recommended. For this reason, more than one method was used in the study.

DIF Detection Methods

Along with the traditional DIF detection methods utilized in the study, this section gives a brief explanation of the DIF detection methods employed in CDMs.

Methods based on CTTs

The Mantel-Haenszel (MH) is a non-parametric uniform DIF determination technique, although being an $\chi 2$ technique suitable for items scored as 1- or for correct/incorrect responses. When DIF has established an advantage across the ability distribution in favor of only one group, this is known as uniform DIF (Swaminathan & Rogers, 1990). This technique splits students into focal and reference groups, classifying the observed scores into several categories. The students with the same test scores will also have the same ability level after the comparison of the scores of the individuals in the groups in terms of their probability of answering the items correctly according to these categories.

In the first step, the method calculates the likelihood ratio for ability levels. As stated by Camilli and Shepard (2004), in order to facilitate the interpretation of these values, the standardized ΔMH value is obtained by taking the natural logarithm of the odds ratios obtained by dividing the odds values of the focal and reference groups. When the ΔMH value is compared to determine whether it is positive or negative, a "+" value indicates that the focal group is superior. In contrast, a "-" value indicates that the reference group is superior. Below are the values for the size of the ΔMH effect, according to Dorans and Holland (1993): |ΔMH|<1 No DIF (Level A); 1<|ΔMH|<1.5 moderate DIF (Level B); 1.5>|ΔMH| and large DIF (Level C).

In Logistic Regression (LR), which is one of the methods that can be used for both DIF types, it was stated that the scores of the items were predicted by group membership and total score. (Zumbo, 1999). While the item is the dependent variable in the technique, the independent variables are the reference and focal groups, and the significance of the effect of two different groups on the item scores is examined. To determine the DIF magnitude for this technique, Zumbo and Thomas (1996) proposed an effect size measure (ΔR^2), widely used in the literature. When the values are examined, the acceptable limit values are ΔR^2<0.13 (Negligible DIF level), 0.13<ΔR^2<0.26 (Moderate DIF level), and ΔR^2>0.26 (Large DIF level).

Methods based on IRTs

Lord's $\chi 2$ is a technique used for both types of DIF. In this technique, item parameters (a_i - item slope parameter and b_{ij} - the item threshold parameter) for the reference and focal groups are calculated separately for each group. The differences in the parameters are controlled according to the IRT model, and the response status of the focal and reference groups to the relevant item is taken into account (Camilli & Shepard, 1994). DIF analysis is used to see if these parameters are the same. It may also be used to test a null hypothesis: "There is no difference between the item parameters between the focal and reference groups." The presence of DIF and the size of the existing effect can be examined by looking at the p and $\chi 2$ values obtained (Hasançebi, 2021).

DIF is connected with the existence or absence of the area between the item characteristic curves (ICCs) in several methods in the literature. Lord (1980) stated that DIFs might occur because one of the two groups with the same ability level at all θ levels has a higher chance of answering the item correctly than the other group. When the ICCs of the two groups intersect, he also pointed out that the DIF for the items becomes complicated. Raju (1988), on the other hand, suggested formulas for calculating the area between the estimated ICCs for the focal and reference groups for one-, two-, and three-parameter models (Camilli & Shepard, 1994). One of these formulas, known as Raju's unsigned area measures technique, is frequently used in the literature to determine both uniform and non-uniform DIFs. The presence of the area between the ICCs obtained for the focal and reference groups was linked to DIF in this technique. When there is no specified area between two ICCs, it means that the item does not have a DIF.

The technique of Raju's unsigned area is popular for determining uniform and non-uniform DIFs in the literature. For one, two, and three-parameter models, Raju (1988) provided methods for determining the

area between the item characteristic curves (ICC) generated for the focal and reference groups (Camilli & Shepard, 1994).

Methods based on CDMs

When the literature is examined, it is stated that the Wald test detects DIF in DINA by using multivariate hypothesis tests. The Wald test, when the focal and reference groups are taken into account, is based on an alternative hypothesis that at least one of the item parameters is different between these two groups. This technique estimates the attribute distributions and item parameters for the focal and reference groups with separate calibrations. In the second stage, the null hypothesis regarding the item parameters of the two groups is tested (Hou et al., 2014). Ma et al. (2021) proposed a new multi-group CDM (MG G-DINA), which enables the responses from different groups to be modelled at the same time, to improve the Wald test's performance in detecting DIF by explaining that students in different groups can use the same or different attributes and they compared the Wald test based on this model and the LRT in detecting DIF. More than one group is calibrated simultaneously in the Wald test based on this model.

Likelihood-ratio test (LRT) is another DIF detection technique used in the MG G-DINA model. According to the literature, this approach based on IRT can be applied under MG G-DINA without any substantial changes. Uniform DIF occurs in cognitive diagnostic assessments when an item supports one of the groups in all attribute profiles. Otherwise, it indicates that non-uniform DIF is present. More detailed information on DIF detection methods, such as MG G-DINA and Wald test and LRT may be found in the studies of Ma et al. (2021) and Mehrazmay et al. (2021).

Methods

Data and Participants

The sample of this study comes from the 2011 administration of the International Association for the Evaluation of Educational Achievement's (IEA) Trends in International Mathematics and Science Study (TIMSS). TIMSS is an independent, international cooperative of national educational research institutions and governmental research agencies dedicated to improving education (Mullis et al., 2009). The sample of this study consists of 488 8th-grade students (48.57% female) who participated in TIMSS 2011 from Turkey. Turkish students tested on Booklet 2 were selected for DIF analyses in this study.

Structure of the Q-Matrix

A Q-matrix consisting of thirteen attributes and thirty-one items developed by Sen and Arıcan (2015) was used in the study. In order to determine the qualifications, the researchers examined the "common core government standards (CCSS)" used to improve the quality of mathematics education. The attribute list of four content areas accepted by the CCSS in 2010 was considered. In mathematics education, four doctoral students matched the items with these attributes. At least two doctoral students must agree that the item is related to the attributes in the Q matrix and that thirty-one items are related to thirteen attributes in the Q matrix (Sen & Arıcan, 2015).

Data Analysis

This study compares DIF detection methods based on CDMs with those based on CTT and IRT. For this purpose, within the scope of the study, gender was considered as a variable, and the analyses were carried out over "Reference group (R): Male students" and "Focal group (F): Female students". The assumptions in the study were examined before proceeding with DIF analyses based on IRT. The two-parameter logistic model (2PLM), which had a considerably better fit, was used for IRT-based DIF analyses. Before proceeding to DIF analyses based on CDM, similar approaches were performed, and

the reduced models were compared to a saturated model, G-DINA. Table 1 shows the results based on the relative fit indices obtained for the model selection that demonstrated the best fit to the data. Although the exact cutoff values for -2LL, AIC, BIC, CAIC, and SABIC relative fit indices have not been determined in the literature, the values of these indices used in model comparisons should be small.

Table 1

Comparing G-DINA to Reduced Models with Relative Fit Indices

Model	-2LL	AIC	BIC	CAIC	SABIC	χ^2	df	p-value
G-DINA	**14238.5**	**30848.5**	65649.1	73954.1	39289.3			
DINA	14983.7	31489.7	66072.4	74325.4	39877.7	745.2	52	<.001
DINO	15148.6	31654.6	66237.2	74490.2	40042.5	910.0	52	<.001
ACDM	14368.8	30918.8	65593.6	73868.6	39329.1	130.3	30	<.001
LLM	14304.7	30854.7	**65529.5**	**73804.5**	**39265.0**	66.17	30	<.001

G-DINA: *Generalized deterministic, noisy "and" gate,* ***DINA:*** *Deterministic, noisy "and" gate,* ***DINO:*** *Deterministic input, noisy "or" gate,* ***A-CDM:*** *additive CDM,* ***LLM:*** *linear logistic model.*

When the values of -2LL and AIC indices are examined, it is seen that G-DINA fits the data better than DINA, DINO, and ACDM. On the other hand, the BIC, CAIC, and SABIC indices show that the values in LLM are small, and the model fits the data better than G-DINA. The LR (likelihood ratio) test can be used to compare the more complex model (G-DINA) to the reduced model (LLM) in such situations (Ma & de la Torre, 2019b). The null hypothesis (H_0: The reduced model's fit to the data is as good as the more complex model) was tested in the LR test for this purpose, and the findings were reported in the table's "p-value" column. When referring to the table, it is clear that the LR test result is significant. G-DINA fits the data better than LLM, as demonstrated by as well. This study used MG G-DINA, a multi-group comparison extension of G-DINA, to provide diagnostic comparisons of male and female students' mathematics performance in the Wald test and LRT-based DIF analyses for the TIMSS 2011 assessment. Although MG G-DINA can be used for more than two groups, it was applied in this study for two different groups, as it did in Ma et al. (2021). All analyses for CTT, IRT, and CDM were performed in the R software using packages "GDINA" (Ma & de la Torre, 2020), "CDM" (Robitzsch et al., 2014), and "difR" (Magis et al., 2018). In the study, the p-values in determining the DIF for multiple comparisons between different methods were corrected using the Holm method, as Ma et al. (2021) used to control familywise error rates at the nominal level of .05.

Results

Results for CTT-based DIF Detection Methods

Findings were reported according to the MH and LR techniques, which are CTT-based DIF detection methods.

Figure 1 and Table 2 show the findings obtained using the MH technique.

Figure 1 displays DIF in seven items (X2, X3, X4, X14, X18, X20, and X27). Among these items, X18 moves far away from the critical value, while X27 moves slightly away from this value. This situation indicates that the largest DIF effect is in X18, and the lowest DIF effect is in X27. Considering the p values in Table 2, and when the ΔMH values obtained from the MH technique for significant items are examined, it is seen that the findings in Figure 1 support the table, and seven items show DIF.

Figure 1

DIF Results Using the MH Technique

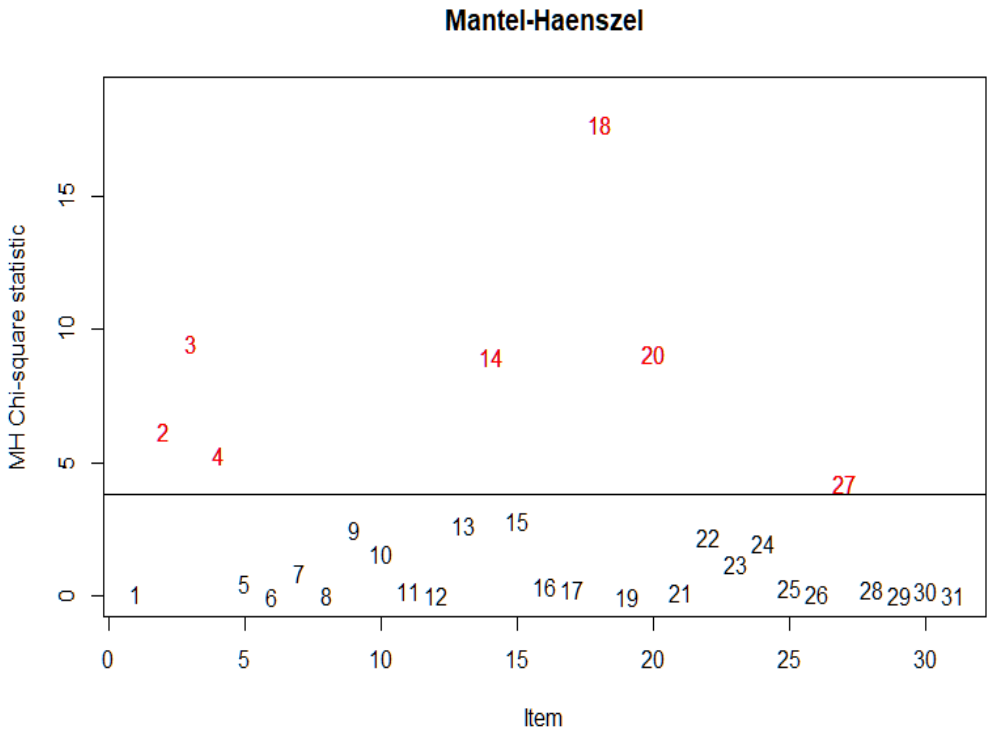

Table 2

DIF Results of the MH Technique

Item	χ^2	*p*	Alpha MH	ΔMH	Effect Size
X1	0.07	.78	0.92	0.19	A
X2	**6.16**	**.01***	0.58	1.26	B
X3	**9.50**	**.00***	2.33	-1.99	C
X4	**5.32**	**.02***	0.55	1.36	B
X5	0.50	.47	0.83	0.41	A
X6	0.00	.94	1.02	-0.05	A
X7	0.90	.34	0.77	0.61	A
X8	0.01	.90	1.14	-0.32	A
X9	2.49	.11	1.56	-1.04	B
X10	1.615	.20	1.41	-0.81	A
X11	0.19	.65	0.88	0.29	A
X12	0.04	.84	0.88	0.27	A
X13	2.66	.10	1.52	-0.99	A
X14	**9.01**	**.00***	0.39	2.17	C
X15	2.80	.09	1.53	-1.00	B
X16	0.35	.55	0.86	0.34	A
X17	0.28	.59	1.17	-0.37	A
X18	**17.71**	**.00***	3.29	-2.80	C
X19	0.00	.93	0.99	0.01	A
X20	**9.10**	**.00***	0.35	2.40	C
X21	0.13	.71	0.79	0.52	A
X22	2.23	.13	1.64	-1.17	B

Table 2

DIF Results of the MH Technique (Continued)

Item	χ^2	*p*	Alpha MH	ΔMH	Effect Size
X23	1.23	.26	0.74	0.69	A
X24	2.00	.15	0.67	0.91	A
X25	0.30	.58	0.85	0.36	A
X26	0.07	.78	1.12	-0.27	A
X27	**4.21**	**.04***	1.57	-1.06	B
X28	0.25	.61	0.88	0.28	A
X29	0.03	.84	0.93	0.16	A
X30	0.20	.65	1.13	-0.29	A
X31	0.02	.86	0.93	0.15	A

Effect size: 0=A; 1.0=B; 1.5=C *'A': Negligible effect; 'B': Moderate effect; 'C': Large Effect*

**p<.05*

Table 2 includes information on DIF's effect size and the magnitude of DIF. Four items (X3, X14, X18, and X20) have a large effect (C level) when the DIF levels of these items are evaluated. Three items (X2, X4, and X27) have a moderate effect (B level). ΔMH values have been examined to see if they were positive or negative, with "+" values favoring the focal group (female) and "-" values favoring the reference group (male). Items X2, X4, X14, and X20 provide an advantage for female students, whereas items X3, X18, and X27 provide an advantage for male students.

Figure 2 and Table 3 present the results obtained using the LR technique.

Figure 2

DIF Results Using the LR Technique

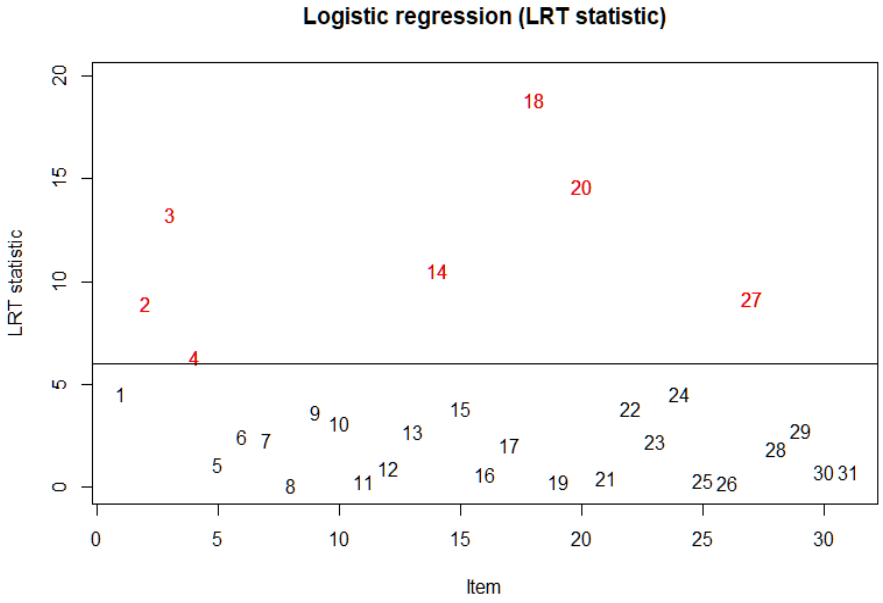

Figure 2 displays DIF in seven items (X2, X3, X4, X14, X18, X20 and X27). Among these items, it is seen that X18 moves far away from the critical value while X4 moves slightly away from it. This situation indicates that the largest DIF effect is in X18, and the lowest DIF effect is in X4. The magnitude of the DIF was determined using the ΔR^2 values in Table 3.

Table 3

DIF Results of the LR Technique

Item	χ^2	*p*	ΔR^2	Effect Size
X1	4.56	.10	0.00	
X2	**8.92**	**.01***	0.06	A
X3	**13.24**	**.00***	0.02	A
X4	**6.33**	**.04***	0.01	A
X5	1.09	.57	0.00	
X6	2.48	.28	0.00	
X7	2.30	.31	0.00	
X8	0.08	.95	0.00	
X9	3.66	.16	0.00	
X10	3.13	.20	0.00	
X11	0.30	0.86	0.00	
X12	0.93	.62	0.00	
X13	2.72	.25	0.00	
X14	**10.56**	**.00***	0.03	A
X15	3.81	.14	0.00	
X16	0.63	.72	0.00	
X17	2.05	.35	0.00	
X18	**18.83**	**.00***	0.03	A
X19	0.28	.86	0.00	
X20	**14.60**	**.00***	0.02	A
X21	0.47	.79	0.00	
X22	3.85	.14	0.00	
X23	2.25	.32	0.00	
X24	4.53	.10	0.00	
X25	0.32	.84	0.00	
X26	0.19	.90	0.00	
X27	**9.16**	**.01***	0.02	A
X28	1.86	.39	0.00	
X29	2.77	.24	0.00	
X30	0.76	.68	0.00	
X31	0.72	.69	0.00	

Effect size: 0.01 = A; 0.13 = B; 0.26 = C

**p<.05,*

'A': Negligible effect; 'B': Moderate effect; 'C': Large effect

When Table 3 is examined, it is seen that seven items display DIF according to the LR technique. The DIF in these items is at the A level and has a negligible effect size, according to Zumbo and Thomas (1996)'s effect size (ΔR^2).

Results for IRT-based DIF Detection Methods

The findings were reported according to the Lord $\chi 2$ and Raju's Unsigned Area Measures Technique, which are IRT-based DIF detection methods, respectively. The findings obtained from the Lord $\chi 2$ technique are presented in Figure 3 and Table 4.

When Figure 3 is examined, it is seen that three red-colored items (X3, X18, and X20) above the threshold value show DIF. Among these items, it is seen that X18 moves far away from the critical value while X20 moves slightly away from it. This indicates that the largest DIF effect is in X18, and the lowest DIF effect is in X20.

Figure 3

DIF Results Using the Lord χ^2 Technique

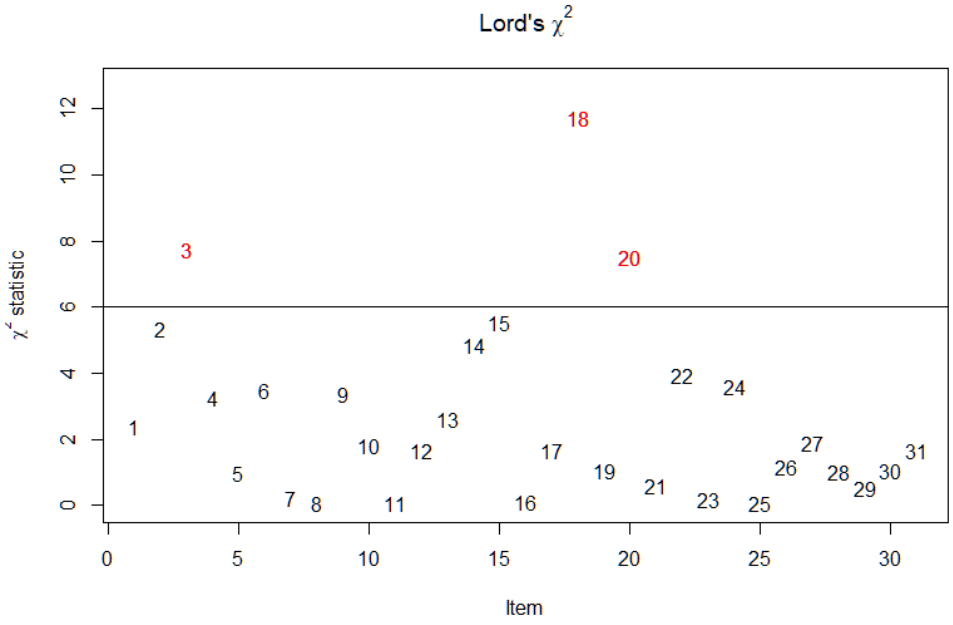

Table 4

DIF Results of the Lord χ^2 Technique

Item	Lord χ^2	*p*	Item	Lord χ^2	*p*
X1	2.38	.30	X17	1.65	.43
X2	5.34	.06	X18	**11.70**	**.00***
X3	**7.74**	**.02***	X19	1.04	.59
X4	3.26	.19	X20	**7.51**	**.02***
X5	0.96	.61	X21	0.60	.73
X6	3.48	.17	X22	3.91	.14
X7	0.22	.89	X23	0.17	.91
X8	0.05	.97	X24	3.58	.16
X9	3.35	.18	X25	0.06	.96
X10	1.80	.40	X26	1.15	.56
X11	0.05	.97	X27	1.88	.38
X12	1.64	.43	X28	1.00	.60
X13	2.59	.27	X29	0.53	.76
X14	4.83	.08	X30	1.04	.59
X15	5.53	.06	X31	1.64	.43
X16	0.10	.94			

**p<.05*

When Table 4 is examined, it is seen that three items show DIF according to the Lord χ^2 technique. Figure 4 and Table 5 display the results of Raju's unmarked area measures technique.

Figure 4

DIF Results Using the Raju's Unsigned Area Measures Technique

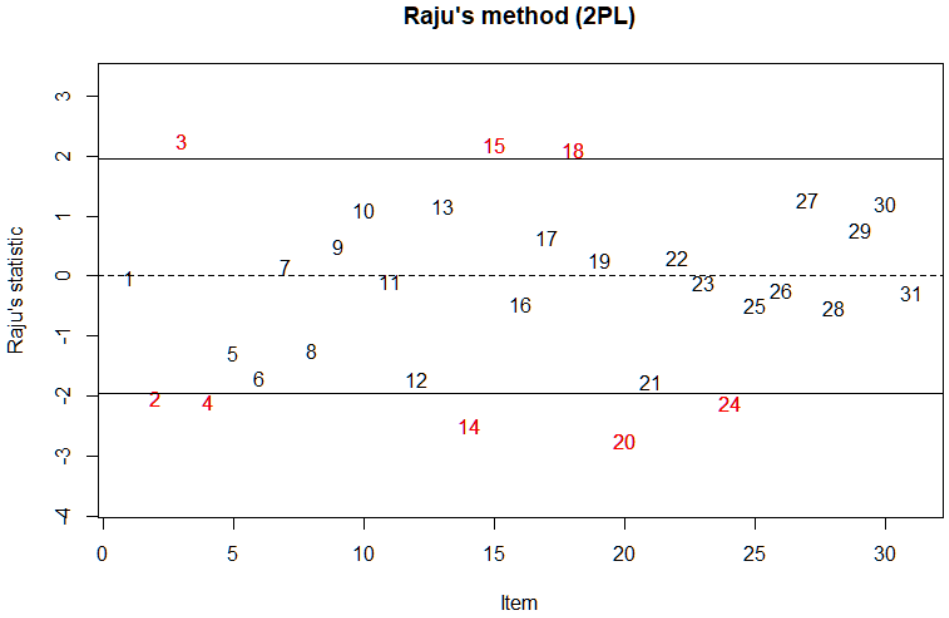

Figure 4 displays eight red-colored items (X2, X3, X4, X14, X15, X18, X20, and X24) above the threshold values that indicate DIF. Among these items, it is seen that X20 moves far away from the critical value while X2 moves slightly away from it. This indicates that the largest DIF effect is in X20, and the lowest DIF effect is in X2.

Table 5

DIF Results of the Raju's Unsigned Area Measures Technique

Item	Raju Statistic	*p*	Item	Raju Statistic	*p*
X1	-0.02	.97	X17	0.64	.52
X2	**-2.03**	**.04***	X18	**2.11**	**.03***
X3	**2.25**	**.02***	X19	0.27	.78
X4	**-2.09**	**.03***	X20	**-2.74**	**.00***
X5	-1.27	.20	X21	-1.75	.07
X6	-1.70	.08	X22	0.31	.75
X7	0.17	.86	X23	-0.10	.91
X8	-1.23	.21	X24	**-2.12**	**.03***
X9	0.49	.62	X25	-0.48	.62
X10	1.10	.26	X26	-0.23	.81
X11	-0.08	.93	X27	1.28	.19
X12	-1.71	.08	X28	-0.52	.60
X13	1.17	.23	X29	0.77	.43
X14	**-2.50**	**.01***	X30	1.21	.22
X15	**2.19**	**.02***	X31	-0.27	.78
X16	-0.46	.64			

**p<.05*

When Table 5 is examined, it is seen that eight items display DIF. Among these items, X2, X4, X14, X20, and X24 provide an advantage in favor of male students. It is seen that items X3, X15, and X18 provide an advantage in favor of female students.

Results for CDM-based DIF Detection Methods

This section used the Wald test and the LRT methods to determine if the test items indicate DIF, and the results are presented in Tables 6 and 7, respectively.

Table 6

DIF Results of The Wald Test

Item	Wald Statistic	Sd	*p*	*d-p*	DIF
X1	2.34	4	.67	1.00	-
X2	2.62	2	.26	1.00	-
X3	3.30	4	.50	1.00	-
X4	5.10	2	.07	1.00	-
X5	0.50	8	.00	1.00	-
X6	2.52	8	.96	1.00	-
X7	2.13	2	.34	1.00	-
X8	8.08	4	.08	1.00	-
X9	7.00	8	.53	1.00	-
X10	9.84	8	.27	1.00	-
X11	0.60	4	.96	1.00	-
X12	2.40	4	.66	1.00	-
X13	1.17	2	.55	1.00	-
X14	1.77	2	.41	1.00	-
X15	8.40	4	.07	1.00	-
X16	0.77	4	.94	1.00	-
X17	0.14	2	.93	1.00	-
X18	7.12	4	.12	1.00	-
X19	5.41	4	.24	1.00	-
X20	2.53	2	.28	1.00	-
X21	0.75	4	.94	1.00	-
X22	0.90	2	.63	1.00	-
X23	2.75	2	.25	1.00	-
X24	0.00	2	.00	1.00	-
X25	1.77	2	.41	1.00	-
X26	13.57	4	.00	.26	-
X27	7.86	4	.09	1.00	-
X28	0.51	2	.77	1.00	-
X29	0.72	2	.69	1.00	-
X30	3.98	4	.40	1.00	-
X31	**17.48**	**4**	**.00**	**.04**	+

'Sd': Degree of freedom; 'd-p': Adjusted p; '-': No DIF; '+': DIF

When the table is examined, it is clear that only one item (X31) displays DIF due to the Wald test. In the Q-matrix Sen and Arıcan (2015) utilized in their studies, this item was related to attributes 3 and 12. The findings obtained with LRT are presented in Table 7.

When the table is examined, it is seen that five items (X9, X10, X13, X20, and X30) show DIF with the LRT technique. Of these items, items X9 and X10 are associated with attributes 8, 9, and 10; item X13 is associated with attribute 13; item X20 is associated with attribute 4, and item X30 is associated with attributes 3 and 13 (Sen & Arıcan, 2015).

Table 7

DIF Results of LRT

	LRT Statistic	Sd	*p*	*d-p*	DIF
X1	4.95	4	.29	1.00	-
X2	5.52	2	.06	1.00	-
X3	9.29	4	.05	1.00	-
X4	-42.66	2	1,00	1.00	-
X5	0.85	8	.99	1.00	-
X6	12.89	8	.11	1.00	-
X7	3.62	2	.16	1.00	-
X8	-23.93	4	1.00	1.00	-
X9	**29.47**	**8**	**.00**	**.00**	+
X10	**53.90**	**8**	**.00**	**.00**	+
X11	-50.46	4	1.00	1.00	-
X12	9.42	4	.05	1.00	-
X13	**14.08**	**2**	**0.00**	**.02**	+
X14	4.59	2	.10	1.00	-
X15	-32.05	4	1.00	1.00	-
X16	11.02	4	.02	.65	-
X17	0.45	2	.79	1.00	-
X18	5.74	4	.21	1.00	-
X19	7.73	4	.10	1.00	-
X20	**33.25**	**2**	**.00**	**.00**	+
X21	10.10	4	.03	.92	-
X22	2.51	2	.28	1.00	-
X23	5.06	2	.08	1.00	-
X24	-0.00	2	1.00	1.00	-
X25	-7.60	2	1.00	1.00	-
X26	4.72	4	.31	1.00	-
X27	13.86	4	.08	.20	-
X28	1.21	2	.54	1.00	-
X29	-2.28	2	1.00	1.00	-
X30	**20.65**	**4**	**.00**	**.01**	+
X31	-23.93	4	.00	1.00	-

'Sd': Degree of freedom; 'd-p': Adjusted p; '-': No DIF; '+': DIF

The probability of having the attributes of interest and the prevalence according to the group's gender variable was investigated to better understand the items with DIF in CDMs and are given in Tables 8 and 9.

Students are assigned to one of the C latent classes using attribute probability. In the Turkey sample, there are 8,192 latent classes for 13 attributes. The prevalence estimate for an attribute is calculated by summing the probability for all relevant latent classes. Table 8 shows that the easiest attribute for male students is N10 (“Understands congruence and similarity using physical models, transparencies, or geometry software.”). About 58% of males have this attribute. The most difficult attributes for males are N1 (“Possesses an understanding of fraction equivalence and ordering; uses equivalent fractions as a strategy to add and subtract fractions”) and N5 (“Reasons about and solves one-variable equations and inequalities; uses properties of operations to generate equivalent expressions.”) because only 31% of males have these attributes.

Table 8

The Prevalence of Attribute by Gender

Attribute	Female	Male
N1	*.33*	*.31*
N2	.46	.40
N3	*.33*	.46
N4	.40	.38
N5	.38	*.31*
N6	.40	.32
N7	.44	.40
N8	.42	.38
N9	.34	.32
N10	.44	**.58**
N11	**.50**	.44
N12	.43	.33
N13	.41	.45

For female students, while the easiest attribute is N11 ("Recognizes perimeter, understands concepts of area, and relates area to multiplication and addition.") which is possessed by 50% of the students, the most difficult attributes are N1 and N3 ("Understands ratio concepts, and uses ratio reasoning to solve problems; finds a percent of a quantity as a rate per 100."), possessed by 33% of the students. In addition, it is seen that female students are more likely to master than male students in the remaining ten attributes except for N10, N3, and N13 ("Investigates chance processes and develops, uses, and evaluates probability models.").

Table 9

Profiles of Attributes of Students by Gender

Latent Class	Female	Male	Latent Class	**Female**	Male
0000000000000	**.14**	**.15**	0010000001000	.00	.04
0000000000100	.05	.04	0010000001100	.00	.04
0000000001000	**.07**	**.07**	0100100100000	.02	.00
0000000001100	.03	.00	0101001100111	.00	.02
0000010000000	.05	.00	1111101111111	.00	.02
0000110000000	.03	.00	1111111110111	.04	.00
0000110001000	.02	.00	**1111111111111**	**.07**	**.08**

When Table 9 is examined, it is seen that 14% of males and 15% of females are in the "0000000000000" latent class. That is, they have not mastered any of the attributes. In the latent class "11111111111111", which represents mastery of all attributes, 7% of females and 8% of males take place. In terms of comparison-based gender, although it is seen that males have a higher rate of mastering all attributes than females, the difference is about 1%. "0000000001000" is another common latent class. When this latent class is investigated, it is observed that only N10 is mastered by 7% of female and male students.

As seen in the findings obtained using MG G-DINA, it is seen that this model provides a diagnostic comparison of the mathematics performance of females and males in the TIMSS 2011 assessment within the scope of cognitive diagnostic assessments.

Comparison of the Methods

In this section, the responses given by male and female students who took the second booklet of the mathematics test in TIMSS 2011 Turkey sample to the items were analyzed to see if the items in the test displayed DIF or not and if the findings were presented in figures and tables according to gender.

The study results based on all methods are presented in Table 10, and comparisons of the methods are made.

Table 10

Comparison of DIF Results of Different Methods

	Traditional Methods					Based-CDM DIF Methods		
	CTT		*IRT*					
Item	*MH*	*LR*	*Lord* χ^2	*Raju*	**DIF**	*Wald*	*LRT*	**DIF**
X1	-	-	-	-	**0/4**	-	-	**0/2**
X2	+	+	-	+	**3/4**	-	-	**0/2**
X3	+	+	+	+	**4/4**	-	-	**0/2**
X4	+	+	-	-	**2/4**	-	-	**0/2**
X5	-	-	-	-	**0/4**	-	-	**0/2**
X6	-	-	-	-	**0/4**	-	-	**0/2**
X7	-	-	-	-	**0/4**	-	-	**0/2**
X8	-	-	-	-	**0/4**	-	-	**0/2**
X9	-	-	-	-	**0/4**	-	+	**1/2**
X10	-	-	-	-	**0/4**	-	+	**1/2**
X11	-	-	-	-	**0/4**	-	-	**0/2**
X12	-	-	-	-	**0/4**	-	-	**0/2**
X13	-	-	-	-	**0/4**	-	+	**1/2**
X14	+	+	-	+	**3/4**	-	-	**0/2**
X15	-	-	-	+	**1/4**	-	-	**0/2**
X16	-	-	-	-	**0/4**	-	-	**0/2**
X17	-	-	-	-	**0/4**	-	-	**0/2**
X18	+	+	+	+	**4/4**	-	-	**0/2**
X19	-	-	-	-	**0/4**	-	-	**0/2**
X20	+	+	+	+	**4/4**	-	+	**1/2**
X21	-	-	-	-	**0/4**	-	-	**0/2**
X22	-	-	-	+	**1/4**	-	-	**0/2**
X23	-	-	-	-	**0/4**	-	-	**0/2**
X24	-	-	-	-	**0/4**	-	-	**0/2**
X25	-	-	-	-	**0/4**	-	-	**0/2**
X26	-	-	-	-	**0/4**	-	-	**0/2**
X27	+	+	-	-	**2/4**	-	-	**0/2**
X28	-	-	-	-	**0/4**	-	-	**0/2**
X29	-	-	-	-	**0/4**	-	-	**0/2**
X30	-	-	-	-	**0/4**	-	+	**1/2**
X31	-	-	-	-	**0/5**	+	-	**1/2**

'-' No DIF; '+' DIF

As table 10 illustrates, when the MH and LR methods from CTT-based methods are compared, it is observed that both methods exhibit DIF for the same items (seven items). When the Lord's χ^2 (three items) and Raju's unsigned area measures (eight items) IRT approaches are compared, the X3, X18, and X20 items in both methods indicate DIF. Five more items indicate DIF using Raju's unmarked area measures technique. Besides, the technique marks the most DIF items among the six methods. Although there are differences between the traditional methods as a whole, it has been observed that X3, X18, and X20 items indicate DIF according to these traditional methods.

When CDM-based DIF detection methods are compared, the Wald test only indicates DIF in one item (X31), while the LRT indicates DIF in five items (X9, X10, X13, X20, and X30). In addition, three

items (X3, X18, and X20) that indicate DIF in conventional methods are investigated with CDM-based methods, and only item X20 indicates DIF with LRT. The items labelled as having DIF via LRT and Wald tests are totally different.

Discussion

Many psychometric questions regarding detecting DIFs in CDMs still exist. Investigating large-scale assessments in the context of DIF by adapting them into CDMs (Terzi & Sen, 2019) may be one of the disregarded questions because looking for meaning in an assessment without making inferences about validity would not give the expected benefit or have the desired influence on educational policies. The invariance of the parameters of the items in the TIMSS 2011 8th-grade mathematics test was controlled by comparing DIF determination methods based on CTT, IRT, and CDM to ensure the correct use of CDMs by performing a retrofitting study. The compatibility of the methods with each other was evaluated.

All assessments must be fair for students with different characteristics (ethnicity, social, and gender). Because DIF analyses are important in affecting groups' inferences from test items (Hou et al., 2014), the DIF effect has been determined using the gender variable as a variable for six different methods. As a result, the researchers' interest in DIF determinations in the test questions is expected to contribute to the validity of diagnostic assessments as an item with DIF can be a potential item for bias. For this purpose, MG G-DINA, which is one of the multi-group models used within the scope of cognitive diagnostic assessments and takes into account sample heterogeneity, was used. Within the scope of this model, this study was considered necessary in order to evaluate the performances of the Wald test and LRT methods, which are relatively newer than traditional methods, on real data.

Thirty-one items differed for both traditional methods and the methods within the scope of cognitive diagnostic assessments. When CTT-based MH and LR methods were compared, it was determined that the same items displayed DIF in both. When Lord's χ^2 and Raju's unsigned area measurements methods, both based on IRT, are compared, the items X3, X18, and X20 indicate DIF in both. DIF was also identified in five more items using Raju's technique of unsigned area measures. The findings show that CTT and IRT methods provide nearly identical outcomes in their own right. This situation supports the findings of previous studies (Kan et al.,2013; Odabas, 2016). Cokluk et al. (2016) stated that both CTT and IRT, produced with different methods on their own merits, are mostly consistent. When CDM-based DIF detection methods are compared in themselves, the Wald test detects DIF in only one item, whereas the LRT technique detects DIF in five. Furthermore, whereas the Raju Unmarked Area Measures technique in IRT had the largest DIF items, the Wald test technique developed for CDMs had the lowest DIF items. Only item X20 displays DIF with the LRT technique when the performances of three items that display DIF in traditional methods are examined using CDM-based methods. The items labelled as DIF by LRT and Wald tests are completely different. Odabas (2016), within the scope of his research, obtained a wide range of items labelled as DIF as a result of the analyzes performed under CDM under different conditions. So that comparisons should be made with the use of more than one technique for DIF studies in CDM.

In order to better comprehend items with DIF in CDMs, the prevalence and possibility of attributes were investigated in this study. The difference between the two groups is approximately 1% for the two most prevalent latent classes ("00000000000000", "11111111111111"), even though males exhibit greater rates of non-mastery and mastery of all attributes than females.

The findings indicated that methods not based on cognitive diagnosis models display DIF more than others. In contrast, the Wald test and LRT methods based on cognitive diagnosis models have fewer items with DIF. There may be several explanations for this situation. The first is that the test used was not developed within the scope of cognitive diagnostic assessments (Ravand & Baghei, 2019). Since determining the qualifications before the test development and defining the Q-matrix by developing the items related to these properties are the most important points of this evaluation approach, the test's psychometric properties may not have been fully determined due to the deficiencies experienced at this

point. However, as stated by the researchers, considering that the development and use of CDM-based tests are not easy and that the negative situations that may occur in ensuring the validity of the Q-matrix are taken into account, it is seen that many CDM applications are adapted to test data developed with non-CDM-based approaches in large-scale data (Gierl et al., 2010). Similar to this study, Odabas (2016) also developed and used a Q matrix prepared later for a previously developed exam in his research. In this process, the researcher stated that the interaction of matter and property in the Q matrix remained within certain limits. Despite this limitation, as stated by the researcher, it is thought that the preparation of the Q matrix and then the development of the exams will be effective in DIF studies as well as parameter estimation and classification accuracy within the scope of CDM. A second possible situation is that LRT may be sensitive to sample size in rejecting the hypothesis "There is no DIF in the relevant item." Mehrazmay et al. (2021) investigated the sensitivity of LRT to sample size and observed that the number of items with DIF increased when different sample sizes were examined. In their study, Ma et al. (2021) found that item discrimination had a significant impact on DIF determination and that Type I error rates in LRT increased when items had low discrimination. They also underlined that the Wald test tended to be conservative when the sample size was small and the item discrimination was high. Liu et al. (2019) reported that as the number of items with DIF increased, the power of MH and LRT methods decreased. Svetina et al. (2018) noted the difficulties with Q-matrix definitions affected the MH, Wald test, and LRT. These findings could explain inconsistencies in the methods utilized in terms of cognitive diagnostic assessments.

When evaluating the consistency of these methods, it is essential to remember that as the number of DIF items in the test increases, the meanings inferred from the scores decrease, raising questions about the validity of the results. As a result, additional research into these new methodologies is required, particularly in cognitive diagnostic assessments. It would be more effective to look into the contributions of these methods to the tests that have been developed, especially when considering the CDM-related test development processes. In this study, DIF analyses were performed, as in Milewski and Baron (2002), to determine the test's psychometric properties within the framework of CDMs rather than the source of bias. In addition, the results were compared with different DIF determination methods from traditional methods, and their compatibility was examined. DIF is not a direct indicator of bias. Due to the abilities of these subgroups, the items may have an actual effect. The source of the difference should be investigated before making a biased decision. Researchers can take the study further and make more comprehensive determinations about item bias in test structure, scope, and subgroups (Dorans & Holland, 1993) if they would like to. As a limitation of this study, the attribute structure of the DIF items was not examined. Future researchers should consider associating the structure of complexity of items with DIF. In addition, DIF analyses were based on only the gender variable. Researchers can also perform studies utilizing various variables.

Declarations

Author Contribution: Büşra EREN-Literature Review, Writing and Critical Review. Tuba GÜNDÜZ-Materials, Data Collection and Processing Analysis, Interpretation. Şeref TAN-Conception, Design and Supervision.

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

Ethical Approval: This research study complies with research publishing ethics. Secondary data were used in this study. Therefore, ethical approval is not required.

References

Akbay, L. (2021). Impact of retrofitting and item ordering on DIF. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(2), 212-225. https://doi.org/10.21031/epod.886920

Asil, M., & Gelbal, S. (2012). Cross-cultural equivalence of the PISA student questionnaire. *Education and Science*, *37*(166), 236-249. https://eb.ted.org.tr/index.php/EB/article/view/1501

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.*Sage.

Cokluk, O., Gul, E., & Dogan-Gul, Ç. (2016). Examining differential item functions of different item ordered test forms according to item difficulty levels. *Educational Sciences: Theory and Practice, 16*(1), 319-330. http://dx.doi.org/10.12738/estp.2016.1.0329

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: development and applications. *Journal of Educational Measurement, 45*, 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253-273. https://doi.org/10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333-353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*(2), 89-97. https://doi.org/10 .1016/j.pse.2014.11.001

DiBello, L. V., & Stout, W. (2007). IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44*, 285-291. https://doi.org/10.1111/j.1745-3984.2007.00039.x

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Earlbaum. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing, 1*, 318-341. https://doi.org/10.1080/15305058.2010.509554

Hasancebi, B. (2021). *Farklı ölçek tiplerinde değişen madde fonksiyonunun belirlenmesi ve yöntemlerin karşılaştırılması üzerine bir çalışma* [A study on determination of item response function in different scale types and comparison of methods] *(Thesis No.687568)* [Doctoral dissertation, Karadeniz Teknik University]. Council of Higher Education Thesis. https://tez.yok.gov.tr/UlusalTezMerkezi/

Hou, L., de la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1), 98-125. https://doi.org/10.1111/jedm.12036

Hou, L., Terzi R., & de la Torre, J. (2020). Wald test formulations in DIF detection of CDM data with the proportional reasoning test. *International Journal of Assessment Tools in Education, 7*(2), 145-158. https://doi.org/10.21449/ijate.689752

Kan, A., Sünbül, Ö., & Ömür, S. (2013). Examination of the item functions of the 6th - 8th grade exams subtests according to various methods. *Mersin University Journal of the Faculty of Education, 9*(2), 207-222. https://dergipark.org.tr/tr/download/article-file/160893

Kang, C., Yang, Y., & Zeng, P. (2018). Q-Matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement, 43*(527-542). https://doi.org/10.1177/0146621618813104

Lee, S., Han, S., & Choi, S. W. (2021). DIF detection with zero-inflation under the factor mixture modeling framework. *Educational and Psychological Measurement, 1*(21). https://doi.org/10.1177/00131644211028995

Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning* [Unpublished doctoral dissertation]. The University of Georgia.

Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology, 10*, 11-37. https://doi.org/ 10.3389/fpsyg.2019.01137

Lord F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge. https://doi.org/10.4324/9780203056615

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*, 1–26. https://doi.org/10.18637/jss.v093.i14

Ma, W., & de la Torre, J. (2019b). *GDINA: The generalized DINA model framework*. R package version (2.7.3). Retrieved from https://CRAN.R-project.org/package=GDINA

Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement, 45*(1), 37-53. https://doi.org/10.1177/0146621620965745

Magis, D., Beland, S., & Raiche, G. (2018). *difR: collection of methods to detect dichotomous differential item functioning (DIF)* (Version 5.0). https://CRAN.R-project.org/package=difR

Mantel, N. & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute, 22*, 719- 748. https://doi.org/10.1093/jnci/22.4.719

Mehrazmay, R., Ghonsooly, B., & de la Torre, J. (2021) Detecting differential item functioning using cognitive diagnosis models: Applications of the wald test and likelihood ratio test in a university entrance

examination, *Applied Measurement in Education*, *34*(4), 262-284. https://doi.org/10.1080/08957347.2021.1987906

Milewski, G. B., & Baron, P. A. (2002, April, 2-4*). Extending DIF methods to inform aggregate reports on cognitive skills*. [Conference presentation]. The Annual Meeting of the National Council on Measurement in Education, New Orleans, LA. https://files.eric.ed.gov/fulltext/ED466712.pdf

Mullis, I. V.S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y. & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Odabas, M. (2016). *Değişen madde fonksiyonunu belirlemede DINA modelde işaretli alan indeksi, standardizasyon, ve lojistik regresyon tekniklerinin karşılaştırılması* [*The comparison of DINA model signed difference index, standardization and logistic regression techniques for detecting differential item functioning] (Thesis No.446894)* [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis. https://tez.yok.gov.tr/UlusalTezMerkezi/

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, *44*, 267–281. https://doi.org/10.1177/0146621619858675

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://link.springer.com/article/10.1007/BF02294403

Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24-56. https://doi.org/10.1080/15305058.2019.1588278

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2014). *CDM: Cognitive Diagnosis Modeling* (Version 3.12). https://CRAN.R-project.org/package=difR

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*.Guilford.

Sen, S., & Arıcan, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2), 238-253. https://doi.org/10.21031/epod.65266

Svetina, D., Feng, Y., Paulsen, J., Valdivia, M., Valdivia, A., & Dai, S. (2018). Examining DIF in the context of CDMs when the Q-matrix is misspecified. *Frontiers in Psychology*, *9*(696). https://doi.org/10.3389/fpsyg.2018.00696

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Terzi, R., & Sen, S. (2019). A nondiagnostic assessment for diagnostic purposes: Q-matrix validation and item based model fit evaluation for the TIMSS 2011 assessment. *SAGE Open*, 9, 1–11. https://doi.org/10.1177/2158244019832684

Zhang, W. (2006). *Detecting differential item functioning using the DINA model* (Unpublished doctoral dissertation]. University of North Carolina at Greensboro.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223-233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., & Thomas, D. R. (1996, October). *A measure of DIF effect size using logistic regression procedures* [Conference presentation]. The National Board of Medical Examiners, Philadelphia. https://scholar.google.com/scholar?cluster=15614527111689986107&hl=tr&lr=lang_tr&as_sdt=2005&sciodt=0,5&as_ylo=20