Kastamonu Education Journal, 2023, Vol. 31, No:4, 680-694
doi: 10.24106/KEFDERGI-2023-0078

| Research Article / Araştırma Makalesi |

The Investigation of The Differential Item Functions of The 2012 High School Entrance Exam Mathematics Test Based On The G-DINA Model

2012 SBS Matematik Alt Testinin G-DINA Modeli Kullanılarak Model Parametrelerinin ve Değişen Madde Fonksiyonlarının İncelenmesi

Sinem ŞENFERAH[1], **Mahmut Sami KOYUNCU**[2]

Keywords

1. HSEE
2. Mathematics
3. Cognitive diagnostic models
4. G-DINA
5. Differential item functioning

Anahtar Kelimeler

1. SBS
2. Matematik
3. Bilişsel tanı modelleri
4. G-DINA
5. Değişen madde fonksiyonu

Received/Başvuru Tarihi
06.08.2023

Accepted / Kabul Tarihi
25.10.2023

Abstract

Purpose: This study aims to determine the item parameters and estimations of student attribute profiles related to the 2012 8th Grade HSEE Mathematics subtest item responses using the G-DINA model and examine whether the items show DIF according to the gender variable.

Design/Methodology/Approach: The study was conducted on the data of 1.063.570 students who took the 2012 HSEE Mathematics subtest. The analysis was carried out on the target population data to avoid sampling error. This study is a descriptive, survey type study.

Findings: When the model fit indices and comparison results were examined, it was concluded that the model that best explained the behavior of students responding to the items in the 20-item form of the 2012 HSEE Mathematics subtest was G-DINA. When the G-DINA model parameters were examined, it was found that the $\alpha_1 = [0000]$ attribute profile (with 61%) in which none of the four defined attributes were found in the student, and the $\alpha_{16} = [1111]$ attribute profile (with 17%) in which all four attributes were present in the student (17%) were the most common attribute profiles. As a result of DIF analysis within the scope of CDM, and it was identified that item 4 showed a significant uniform DIF in favor of female students while item 19 showed a uniform DIF in favor of male students at moderate level.

Highlights: It is thought that DIF analyses within the framework of cognitive diagnostic models can provide a statistical basis for item bias decisions.

Öz

Çalışmanın amacı: G-DINA model kullanarak 2012 SBS 8. Sınıf Matematik alt testi madde cevaplarına ilişkin madde parametreleri ve öğrenci nitelik profili kestirimlerinin belirlenmesini ve maddelerin cinsiyet değişkenine göre DMF gösterip göstermediğinin incelenmesidir.

Materyal ve Yöntem: Çalışma, 2012 yılı SBS Matematik alt testini alan 1.063.570 öğrenci verisi üzerinde yürütülmüştür. Analizler, örnekleme hatasının önüne geçmek amacıyla evren verisi üzerinde gerçekleştirilmiştir. Betimsel düzeyde, tarama türünde bir araştırmadır.

Bulgular: Model uyumu indeksleri ve karşılaştırma sonuçları incelendiğinde, öğrencilerin 2012 SBS Matematik alt testinin 20 maddelik formunda yer alan maddelere cevap verme davranışlarını en iyi açıklayan modelin GDINA olduğu sonucuna ulaşılmıştır. GDINA model parametreleri incelendiğinde, tanımlanan dört nitelikten hiçbirinin öğrencide bulunmadığı $\alpha_1 = [0000]$ nitelik profili (%61) ile dört niteliğin de öğrencide bulunduğu $\alpha_{16} = [1111]$ nitelik profilinin (%17) en çok rastlanan nitelik profili olduğu belirlenmiştir. BTM kapsamında DMF analizleri sonucunda, 4. maddenin önemli düzeyde kız öğrenciler lehine tek biçimli DMF gösterdiği; 19. maddenin ise orta düzeyde erkek öğrenciler lehine tek biçimli DMF gösterdiği bulunmuştur.

Önemli Vurgular: Bilişsel tanı modelleri çerçevesinde DMF analizleri ile madde yanlılığı kararlarına istatistiksel bir zemin oluşturulabileceği düşünülmektedir.

[1] Kastamonu University, Education Faculty, Educational Sciences, Kastamonu, TÜRKİYE; https://orcid.org/0000-0001-7932-7644

[2] **Corresponded Author**, Afyon Kocatepe University, Educational Sciences, Afyonkarahisar, TÜRKİYE; https://orcid.org/0000-0002-6651-4851

Citation/Alıntı: Şenferah, S., & Koyuncu, M. S. (2023). The Investigation of The Differential Item Functions of The 2012 High School Entrance Exam Mathematics Test Based On The G-DINA Model, *Kastamonu Education Journal, 31*(4), 680-694. doi: 10.24106/KEFDERGI-2023-0078

INTRODUCTION

Central examinations conducted by the Ministry of National Education (MoNE) are decisive in the secondary education placement process in Türkiye. In our country, the secondary education placement exams held at the national level by the MoNE have differed in terms of method and content in various periods. Student selection and placement was carried out centrally, sometimes by applying a single exam and sometimes more than one exam. Although it was expressed with different names (OKS, SBS, TEOG, LGS, and so on.) in different periods, the common purpose of the exams is to select and place students for high schools. In our country, the exam conducted by MoNE between 2007-2013 to place students in secondary education is the High School Entrance Exam (HSEE). Ensuring the validity of large-scale exams such as HSEE, which had special purposes such as selection and placement, and whose results could directly affect the lives of students, accordingly, was of great importance in terms of the accuracy of the decisions taken.

Determining whether the test items in the exams provide an advantage to any of the subgroups due to the characteristics of the test taker groups that are not related to the measured attribute by differential item function (DIF) analysis is of great significance as evidence of the validity of the test scores and the decisions taken accordingly. DIF analysis, which has become a part of item analysis in recent years, is routinely performed to ensure the validity and fairness of test scores. Although many DIF methods have been developed within the scope of Classical Test Theory (CTT) and Item Response Theory (IRT), the applications of these methods in cognitive diagnosis models (CDM) are quite limited.

Cognitive diagnosis models are latent classroom models developed to assess whether students have interrelated but distinguishable latent attributes (de la Torre, 2011; Hou et al., 2014; Haagenars & McCutcheon, 2002). The fact that cognitive diagnosis models are more related to classroom teaching and learning processes and that they provide more diagnostic information have caused them to be considered more as a psychometric research topic (Rupp & Templin, 2008; de la Torre, 2011; de la Torre & Douglas, 2004; Embretson, 1997; Junker & Sijtsma, 2001; Tatsuoka, 1985). Cognitive diagnostic models have been developed as an alternative to one-dimensional item response models to determine whether the respondent has the multi-component skills required to answer test items correctly (de la Torre, 2009).

IRT models only provide a single score that assesses the overall ability level of the respondent. However, the CDM analysis provides a profile that shows what attributes, skills, or knowledge each respondent has. These profiles contain important information that can have an impact on learning and teaching. Instead of just giving a total score, CDM shows the attributes, strengths, and weaknesses of individuals and the reasons for failure in detail. This provides valuable guidance for better guiding learning processes and improving instructional design. Traditional IRT methods examine whether the probability of answering the item correctly differs between respondents in different subgroups who are at the same ability level or get the same total score (Hambleton et al., 1991; Zumbo, 1999). DIF within the scope of CDM, on the other hand, evaluates the differentiation of the probability of answering the item correctly among respondents who are in different groups and have the same latent attribute profile. DIF assessments under CDM provide evidence of the invariance of the item-attribute interaction between groups (Hou et al., 2014). According to Hou et al. (2014), the presence of items with DIF in terms of CDM may lead to item parameters and latent trait profile estimations that are not valid for each group. Therefore, DIF analyzes are necessary to determine parameter and structure invariance (Zumbo, 2007). Since item parameters represent interactions between attributes and items, DIF analysis sheds light on whether the attribute-item interaction is invariant between comparison groups. Group membership, which can affect how items are perceived and resolved, may act as a variable leading to DIF. For this reason, controlling the invariance between groups with DIF analysis is considered as the necessary first step of CDM applications (Hou et al., 2014).

In the context of CDM, DIF can be represented by $\varDelta_{j\alpha_l} = P\left(X_j = 1\middle|\alpha_l\right)_F - P\left(X_j = 1\middle|\alpha_l\right)_R$. Here $\varDelta_{j\alpha_l}$ indicates item j. in DIF for respondents with latent attribute profile; α_l the probability of success in the item j. for respondents with latent attribute profile. A value of $\varDelta_{j\alpha_l} > 0$ indicates that the item shows DIF in favor of the focus group, while a value of $\varDelta_{j\alpha_l} < 0$ indicates that the item shows DIF in favor of the reference group. The fact that $\varDelta_{j\alpha_l} = 0$ for all attribute profiles is interpreted as the case where the item does not show DIF.

Similar to DIF under IRT, uniform and non-uniform DIF can be identified in cognitive diagnostic models. Regardless of the latent attribute profile, if the probability of answering an item correctly is consistently higher or lower for a group, in other words, if $\varDelta_{j\alpha_l}$ is either positive or negative in all latent attribute profiles, the item shows uniform DIF. While the probability of answering the item correctly is lower in some latent trait profiles for a group, and while it is higher in some other latent profiles for the same group, non-uniform DIF is the case there. Namely, in non-uniform DIF, the sign of $\varDelta_{j\alpha_l}$changes depending on the latent attribute profiles (Hou et al., 2014).

In the current study, the aim is to determine whether the 2012 HSEE 8th Grade Mathematics subtest items based on the G-DINA model, which is one of the cognitive diagnosis models, show DIF according to the gender variable, and to identify the model parameters and student qualification attributes. It is thought that this study will contribute to the relevant literature due to the limited number of studies in which DIF studies are handled within the scope of CDM (Milewski & Baron, 2002; Zhang, 2006; Li, 2008; Hou et al., 2014; Li & Wang, 2015). When the relevant literature is analyzed, it is seen that studies using cognitive diagnosis models are generally carried out on simulation data (de la Torre & Douglas, 2004; Zhang, 2006; Hou et al., 2014; Li & Wang, 2015; Ömür Sünbül & Kan, 2015). In the current study, it is thought that large-scale HSEE real data at the national level will make important contributions to the field in terms of seeing how the G-DINA model will produce results, providing empirical information about learning and teaching processes, and providing evidence for the validity of test scores.

METHOD/MATERIALS

Research Design

This study is a descriptive, survey type study as it aims to determine whether the 2012 HSEE 8th Grade Mathematics subtest items show DIF according to the gender variable, and to estimate model parameters and student attribute profiles using the G-DINA model. The purpose of survey research is generally to make a description by taking a picture of an existing situation related to the research topic (Büyüköztürk et al., 2014). Survey studies are studies conducted on larger samples compared to other studies in which certain characteristics of a group (e.g., ability, attitude, belief, and knowledge) are described (Fraenkel & Wallen, 2012).

Population

The population of the study consists of 1.075.546 students who participated in the 8th Grade High School Entrance Exam in 2012. The study was carried out on the data of a total of 1.063.570 students, excluding the students whose gender information in the 2012 HSEE Mathematics subtest could not be accessed from the analysis. The students that were removed from the data set constituted approximately 1% of the data set. The analysis was carried out on the whole target population data to avoid sampling error. The distribution of students by gender for the mathematics subtest is provided in Table 1.

Table 1. The Distribution of the students according to gender

Gender	N	%
Female	523.939	50,7
Male	539.631	49,3
Total	1.063.570	100

When Table 1 is examined, it is seen that 523.939 (50,7%) of them are females and 539.631 (49,3%) of them are males.

Data Collection

Depending on the purpose of the study, the data collection process includes obtaining the item response data for the Transition to Secondary Education System (which is done via HSEE) 2012 8th Grade Mathematics subtest and creating the Q-matrix that defines the relationships between the items and the attributes. The data of the 8th grade mathematics subtest of the HSEE held in 2012 were provided by the Ministry of National Education, General Directorate of Assessment and Examination Services. As for the Q matrix, it was determined through the focus group discussion with the participation of 7 field experts.

According to the guide of the Transition to Secondary Education System (TSES) Placement test, 20 multiple choice mathematics questions with 4 options were directed to the students in the 8th grade. The descriptive statistics regarding the mathematics subtest used in the study were provided in Table 2.

Table 2. The descriptive statistics of the mathematics subtest

Statistics	Mathematics Subtest
Number of Item	20
Number of Students	1.063.570
Min. Score	0
Max. Score	20
Mean	7.17
Variance	22.60
Standard Deviation	4.75
Skewness	0.94
Kurtosis	0.07
KR-20	0.86
Standard Error of Measurement	1.79
Mean Difficulty ($\overline{p}$)	0.36
Mean Discrimination (r_{pb})	0.51

When Table 2 is examined, it is seen that the reliability coefficient (KR-20) for the mathematics subtest is 0.86. That the reliability coefficient calculated is higher than 0.70 in terms of internal consistency is generally considered sufficient for the reliability of the test scores (Büyüköztürk, 2012). In addition, a mean difficulty value of ($\overline{P}$) 0.36 indicates that the students who took the test answered 7.17 of 20 items on average in the test, and that skewness and kurtosis coefficients are positive indicates that the score distributions of the mathematics subtest are skewed to the right, which means that there is an accumulation in low scores.

The Formation of the Q-matrix

The quality of diagnostic evaluations is influenced by the correct identification of the attributes underlying test performance. It is stated that various sources such as test guides, learning domain theories, item content analysis, analysis of the respondent's test process, and related research results in the literature can be utilized to determine the attributes covered in a test (Embretson, 1991; Leighton & Gierl, 2007). Lee and Sawaki (2009a), on the other hand, stated that when cognitive models of task performance are not available and cognitive diagnostic models are used for non-diagnostic tests, it is a good starting point to brainstorm about possible attributes by examining the test content in detail. Accordingly, in this study considering that the HSEE, which is used for selection and placement purposes, was not developed for diagnostic purposes, the study aimed to determine the possible attributes to be measured with the HSEE mathematics subtest through focus group interviews.

The focus group discussion process, which was carried out to determine the Q-matrix that defines the relationships between items and attributes and shows whether a feature is necessary for an item, was completed in 4 stages. In the focus group interview, firstly, the possible attributes to be measured with the mathematics subtest and the gains associated with the items were examined and the boundaries, main topics and interview questions of the focus group interview were determined. In the second stage, the field experts who were going to participate in the research were identified and invited to the interview. In the third stage, the place and time for the focus group meeting were arranged, the necessary arrangements were made, and the interview was held. As for the last stage, the short notes taken during the interview were analyzed and the results were summarized.

A principle and 7 field experts participated in the focus group meeting held to determine the attributes required for answering the items correctly. Although there are different opinions about how many people the group size should consist of in focus group interviews, ideally 6-8 people are considered sufficient (Yıldırım & Şimşek, 2011). Since the identification of the attributes required for answering the items correctly and the definition of the relationships between the items and the attributes require expertise and experience both in the fields of mathematics education and measurement and evaluation, it was ensured that all experts graduated from mathematics education and held at least a master's degree in mathematics education or measurement and evaluation fields. Detailed information on the training areas of the experts participating in the focus group interview is given in Table 3.

Table 3. The distribution of the experts that took part in the focus group meeting

Educational Background	Total Number	In Total (%)
A teacher at MoNE		
➢ Ph.D. student in Mathematics Education	1	12,5
A research assistant	7	87,5
➢ Ph.D. student in Mathematics Education and MA student in Assessment	1	12,5
➢ MA degree from Primary School Mathematics Education and Ph.D. student in Assessment	2	25
➢ Ph.D. student in Assessment	3	37,5
Ph.D. student in assessment (a principle)	1	12,5
Total	8	100

In the focus group discussion, the aim was to create an environment where the participants could hear the opinions of others and think about their own opinions accordingly. In line with this, the environment was prepared in a round seating arrangement so that the experts could see each other. Care was taken to ensure that the environment was noise-free, and that the conversation was not interrupted. At the beginning of the focus group discussion, an explanation was made to the experts regarding the purpose and scope of the study, and a form containing the mathematics items that were the subject of the research and the pedagogic gains thought to be related to these items was distributed. In addition to this form, the classification of learning domain and cognitive skills covered within the scope of TIMSS 2015 8th Grade Mathematics Framework (Gronmo et al., 2014) was also utilized to examine the item contents, relevant learning areas, and the attributes to be measured with a specific item.

In the focus group interview, the acquisition related to the mathematics subtest items were evaluated in terms of the learning areas included in the Mathematics Lesson (the 6-8th Grades), the Curriculum, and TIMSS 2015 8th Grade Mathematics Framework. Additionally, the cognitive skills to be measured with items were investigated under the classifications of "knowing", "practicing" and "reasoning" determined within the scope of TIMSS 2015 8th Grade Mathematics Framework. However, that the strategies used in item solution could not be identified due to the multiple-choice nature of the test items prevented unearthing the relationship between the item and cognitive attributes to be made with full accuracy. Lee and Sawaki (2009a) emphasize that the features defined in more detail in the Q matrix will provide richer diagnostic information and thus increase the instructional value of the diagnosis, but this will generally produce unreliable and inconsistent results in classifying respondents. Consequently, in the current study evaluating the attributes required for the correct answer of the items within the scope of learning areas was decided as the goal. Accordingly, in the Q-matrix, the relationship of the items with 4 attributes, namely "Numbers", "Geometry and Measurement ", "Algebra", and "Probability and Statistics" were defined. As a result of the focus group discussion, the Q-matrix provided in Table 4, which defines the relationships between the items and the attributes, was created.

Table 4. The Q-matrix

Item	Numbers	Geometry and Measurement	Algebra	Probability and Statistics
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	0	1	0	0
5	0	1	1	0
6	0	1	0	0
7	0	1	0	0
8	1	0	1	0
9	0	1	0	0
10	1	0	0	0
11	0	1	0	0
12	0	1	0	0
13	0	1	0	0
14	0	1	0	0
15	0	0	0	1
16	1	0	0	1
17	0	0	0	1
18	0	1	0	0
19	1	0	1	0
20	0	0	1	0
Total	7	10	4	3

When Table 4 is analyzed, it is observed that the Q-matrix consists of 1 entries that show that the relevant attribute is necessary for the item and 0 entries that show that relevant attribute is unnecessary for the item. Hartz et al. (2002) suggest that at least 3 items must be present for an attribute to obtain reliable diagnostic information. According to the created Q-matrix, it was determined that the attributes in the test were measured with at least 3 (Probability and Statistics) and at most 10 items (Geometry and Measurement).

Data Analysis

This study basically serves two purposes; The first one is to determine the model parameters and latent attribute profiles for the mathematics subtest via the G-DINA model, and the second is to examine whether the mathematics items show DIF by gender using the G-DINA model. In line with these purposes, based on the Q-matrix that defines the relationships between the "Numbers", "Geometry and Measurement", "Algebra", and "Probability and Statistics" attributes and the items, whether the probability of answering the items correctly differs between both the model parameters and the students with the same latent attribute profile was analyzed.

As with other statistical models, the significance and interpretability of the results obtained from cognitive diagnostic models depend on the extent to which model data fit is achieved. Model fit can be determined in two ways, in which the fit of the model to the data is checked (absolute fit) and the model is compared with other models (relative fit). In this study, both absolute fit indices (Mx2, MADcor, MADRESIDCOV, MADQ3, and SMRSR) and also relative fit indices (Loglik, AIC, and BIC) were examined in order to test the overall model fit.

Mx2 (Chen & Thissen, 1997), a global model fit test from absolute fit indices, is the average of χ2 test statistics regarding the independence of item response frequencies across all item pairs. Mx2 represents the mean of the difference between the observed and predicted response frequencies by the model. Ravand (2016) stated that significant differences can be taken as evidence of inter-item dependence and that dependence is expected because respondents use the same cognitive processes to answer the items. If the cognitive diagnosis model fits the data well, the χ2 test statistics is expected to be 0. In this case, the attribute profiles of the respondents, that is, each latent class, will be perfectly predicted from the observed response pattern (Rupp et al., 2010). MADcor (DiBello et al., 2007) is the average of the absolute differences between observed and predicted item pair correlations. The "mean residual covariance (MADRESIDCOV)" (McDonald & Mok, 1995) is the mean of the absolute differences between the observed and reproduced item covariance matrices. The MADQ3 (Yen, 1984) is the average of the absolute values of the Q3 statistics describing the binary correlations for item residuals. The average of the RMSEA values at the item level compares the rates observed and predicted by the model for each response category weighted with the ratio of respondents in latent classes (Lei & Li, 2016). Classification consistency (Pc) and classification accuracy (Pa) express reliability and validity regarding classifying respondents into implicit classes. While Pc is the indicator of the extent to which respondents will be consistently classified into the same latent class when the same test or a parallel form of the test is applied, Pa is an indicator of how well the respondent's classification matches the correct latent class (Ravand, 2016).

In the study, within the scope of the G-DINA model, whether the 2012 HSEE Mathematics subtest items displayed item function variation according to gender was investigated via the Wald test (de la Torre, 2011; Hou et al., 2014). On the other hand, it is stated that Wald statistics (de la Torre & Lee, 2013) significantly determines even negligible DIF effects in large samples (George & Robitzsch, 2014). Hence, the unsigned area measurement based on the unmarked area (UA) originally introduced by Raju (1990) is presented as the DIF effect size. In the literature, it is seen that the effect size classification criteria set forth by Jodoin and Gierl (2001) are used for UA (George & Robitzsch, 2014; Ravand & Robitzsch, 2015). Jodoin and Gierl suggested a critical value of .059 for negligible DIF sizes and .088 to determine medium DIF sizes within the scope of three-parameter IRT models.

Data analysis was performed using the RStudio program "CDM" package version 4.99-11 (Robitzsch et al., 2016). The "CDM" package uses the marginal maximum likelihood method based on the EM (Expectation-Maximization) algorithm to obtain parameter estimations.

FINDINGS

The G-DINA Model Fit of 2012 HSEE Mathematics Test Data

The statistics of the G-DINA model fit of the 2012 HSEE Mathematics Test were provided in Table 5 and the values of classification consistency (P_c) and its accuracy (P_a) are presented in Table 6.

Table 5. The statistics of the G-DINA model fit of 2012 HSEE mathematics test

Model fit indices		*p*
Mx2	22603.7	0
MADcor	0.045	-
MADRESIDCOV	0.009	-
MADQ3	0.032	-
RMSEA	0.045	-
SRMSR	0.057	-

When Table 5 is examined, it is seen that the Mx2 statistic is significant (p < 0). Although the Mx2 statistic as an indicator of fit is not expected to be significant, it is sensitive to even minor model data mismatches. For this reason, it often yields results that indicate the lack of harmony. Consequently, it was stated that SRMSR (the Standardized Root Mean Squared Residual) values should also be reported for the Mx2 statistics. A SRMSR value close to zero indicates better model-data fit (De Ayala, 2009; Maydeu-Olivares & Joe, 2005, 2006; Maydeu-Olivares et al., 2011; Maydeu-Olivares & Joe, 2014). Maydeu-Olivares (2013) states that the SRMSR value should be less than 0.05 for the model to fit well. The fact that the obtained SRMSR value is very close to 0.05 indicates that the G-DINA model fits at a good level. MADcor was found to be 0.045 in the study. In their study, DiBello et al. (2007) considered the MADcor value of 0.049 as the indication of the fact that the cognitive diagnosis model fits well with the data. As seen in Table 5, MADRESIDCOV, MADQ3, and RMSEA values were found to be less than 0.05. Ravand (2016) evaluated MADRESIDCOV, MADQ3, and RMSEA values less than 0.05 as a good model fit. In this respect, it may be suggested that the G-DINA model used in the study fits well with the HSEE 2012 Mathematics subtest data.

Table 6. Classification Consistency (P_c) and Accuracy (P_a) for G-DINA Model

	Classification Consistency (P_c)	**Classification Accuracy (P_a)**
Atribute1	.94	.97
Atribute2	.99	.99
Atribute3	.57	.69
Atribute4	.72	.83
Pattern	.92	.89

When Table 6 is examined, it is seen that the P_a and P_c values for the whole latent class pattern in the study are .89 and .92, respectively. P_c and P_a values give the measure of classification consistency and accuracy regarding whether these attributes are owned by students for each attribute. Except for the third attribute (i.e., Algebra), these values for the attributes were found to be relatively high. Ravand (2016) states that although there is no definitive criterion for Pc and Pa values, C.Ying (2013) recommends the values of .7 and .8 as acceptable classification rates. However, in Cui et al. (2012)'s study on Tatsuoka's (2002) subtraction data in fractions, P_a and P_c values were found to be .68 and .52, respectively. In the light of this information, it may be stated that the validity and reliability of the classification are at an acceptable level.

In addition, the DINA model, which is a "conjunctive" model that requires the student to have all the necessary attributes to be successful in the relevant item, and the G-DINA model, where each attribute's contribution to the probability of answering the item is different, were compared and which CDM model better fitted the data was also examined. Table 7 presents the relative fit statistics and likelihood ratio test results obtained from the 2012 HSEE Mathematics subtest according to the DINA and G-DINA models.

Table 7. The relative fit statistics and likelihood ratio test (LR) according to the DINA and G-DINA models

Model	Nobs	AIC	BIC	Npar	logLik	χ^2	df	*p*
DINA	1063570	22233810	22234464	55	-11116850	72939,27	4	.0
G-DINA	1063570	22160879	22161580	59	-11080380			

When Table 7 is examined, it is seen that there is a decrease in the Loglik, AIC, and BIC values regarding the G-DINA model. This can be interpreted as the fact that the G-DINA model fits the data better than the DINA model. The goodness-of-fit test (Bock & Lieberman, 1970) using LRT χ^2 values for DINA and G-DINA model comparison is seen to be significant (p = .0). Accordingly, the G-DINA model fits the data better than the DINA model. When the general model fit indices and comparison results were examined, it was concluded that G-DINA was the model that best explained the students' behavior of responding to the items in the 20-item form of the 2012 HSEE Mathematics subtest.

The Findings regarding the Parameters of the G-DINA Model

The difficulty levels of the attributions were examined within the scope of the study, and the results are given in Table 8.

Table 8. The probabilities regarding the attributes

Attributes	Attribute probabilities
Numbers	.28
Geometry and Measurement	.20
Algebra	.26
Probability and Statistics	.33

When Table 8 is examined, it is seen that 33% of the students have attributes related to the learning field of "probability and statistics". Accordingly, "probability and statistics" can be expressed as the easiest attribute. This attribute is followed by "numbers", "algebra", and "geometry and measurement" respectively.

In cognitive diagnostic models, respondents are classified into 2^K latent classes. In this study, students are divided into 16 latent classes, as 4 attribute areas are defined within the scope of the HSEE Mathematics subtest. In Table 9, possible student attribute profiles for the first two and the last two latent classes are presented.

Table 9. The probabilities regarding latent classes

Latent Classes	Attribute Profile (α_i)	Class Probabilities	Expected Class Frequencies
1	0000	.614	654049.7
2	1000	.026	27926.2
.			
.			
.			
15	0111	.00009	105.984
16	1111	.168	179120.2

In the current study, it was identified that the highest-class probability belonged to the latent attribute profile α_1= [0000]. According to this, approximately 61% of the students (nearly 654.050 students) took part in the first latent class where they were expected to have none of the four attributes. It was determined that the second highest class probability belonged to the latent attribute profile α_{16}= [1111]. Accordingly, approximately 16.8% of the students (nearly 179.120 students) were expected to have all the attributes. G-DINA parameter estimations for all items in the mathematics subtest are given in Table 10.

Table 10. G-DINA model parameter estimations

	Attribute Profile															
	0	1														
Items	00	10	01	11												
	000	100	010	001	110	101	011	111								
	0000	1000	0100	0010	0001	1100	1010	1001	0110	0101	0011	1110	1101	1011	0111	1111
1	0.19	0.54														
2	0.24	0.95														
3	0.27	0.95														
4	0.58	0.93														
5	0.13	0.00	0.11	0.53												
6	0.16	0.79														
7	0.37	0.73														
8	0.17	0.69	0.63	0.90												
9	0.23	0.87														
10	0.12	0.21														
11	0.15	0.84														
12	0.21	0.88														
13	0.14	0.73														
14	0.11	0.47														
15	0.40	0.92														
16	0.39	0.83	0.92	0.97												
17	0.19	0.79														
18	0.10	0.79														
19	0.27	0.80	0.85	0.96												
20	0.11	0.27														

Table 10 shows the probability of being successful in certain attribute profiles for all items in the mathematics subtest. The pattern of reduced attribute profiles to which the parameter estimates correspond is in the top row of the table. In the G-DINA model, the number of parameters for each item is a function of the number of attributes required for that item $(2^{K_j^*})$. Accordingly, four parameters are estimated for items that require two attributes, and two parameters are estimated for items that require one. When Table 10 is examined, it is observed that four parameters are obtained for items 5, 8, 16, and 19, which require two attributes. The point to be noted here is that the required attributes for these items are not the same, that is, the attributes represented in the reduced feature vector are different. Since items other than the items 5, 8, 16, and 19 require a single attribute, P(0) and P(1) can be interpreted as g and 1-s parameters respectively in the DINA model (de la Torre, 2011). According to this, it may be stated that the students who do not have the attribute of numbers guessed the third item correctly with a probability of 27%, while the students with the attributes in the field of learning numbers answered the item correctly with a probability of 95%. In addition to this, it was determined that none of the items requiring more than one attribute in the mathematics subtest met the "conjunctive" assumption of the DINA model, which means that the same probability of success is obtained in the absence of one of the required attributes.

It is seen that the students who have one of the two required attributes in items 8, 16, and 19 show higher success than the students who do not have any of these attributes. This finding indicates that the achievement in the related items changes depending on the presence of the attributes in the student, and therefore, correct definitions of the attributes are provided in the Q matrix. On the other hand, when the 5th item is taken into consideration, it is seen that the main effects of the "geometry and measurement" and "algebra" attributes required for the item are quite low; however, the interaction of these two attributes were observed to affect the probability of success in the item at a high level. When the 4th, 10th, and 20th items are examined, it is seen that there is not a significant difference between the success probability of the students who have the necessary attributes for the item and the success probability of those who do not. This indicates that other qualities may have been used to answer the item correctly, or that other characteristics unrelated to the test, apart from the attributes to be measured with the item, may have affected the answering process. In Table 11, the attribute pattern and probability of success for items 1 and 8 in the mathematics subtest are given, as well as the required attributes for the item.

Table 11. Attribute pattern and success probabilities for items 1 and 8

Item	Attributes Required for the Item	Attribute Pattern	Probabilities of Success
Item1	α_1	A0	.19
Item1	α_1	A1	.54
Item8	α_1 ve α_3	A00	.17
Item8	α_1 ve α_3	A10	.69
Item8	α_1 ve α_3	A01	.63
Item8	α_1 ve α_3	A11	.90

In Table 11, the second column shows the required attributes defined in the Q-matrix for items 1 and 8, the third column shows the patterns of whether the student has the necessary attributes for the item, and the last column shows the probability of answering the item correctly depending on whether the required attributes are present or not. When Table 11 is analyzed, it is seen that the probability of answering the item correctly (estimation probability) of the students who do not have the attributes related to the "numbers" learning area required for item 1 is .19. In Table 11, it is seen that the students who have the attributes related to the "numbers" learning field are more likely to answer item 1 correctly (54%) than those who do not. In the case that the student has the attributes related to the "numbers" learning domain, the probability of being successful in the item (probability of not making a shift) becomes .19+.54= .73. When Table 11 is examined, it is seen that 2 attributes are required for item 8, namely "numbers" and "algebra". It is seen that students who do not have the 2 attributes required for item 8 answered the item correctly with a probability of 16%. Accordingly, the probability of guessing the 8th item in the mathematics subtest is 16%. If the student has the attributes related to the "numbers" learning field, the probability of being successful in the item is .16+.69=.85. If the student has the attributes related to the "Algebra" learning field, the probability of being successful in the item becomes .16+.63= .79. Accordingly, the presence of the "Numbers" attribute for the 8th item affects the success of the item more than the "Algebra" attribute. As for the interaction of "Numbers" and "Algebra", it increased the probability of answering the item correctly to 90%. However, since it is a conditional probability, it is necessary to pay attention not to add up the probability of answering the item correctly, depending on whether the necessary attributes for any items are present or not (Ravand, 2016).

The Findings regarding DIF Analysis

The results of the DIF analysis within the scope of the G-DINA model for the mathematics subtest obtained using the Wald test are provided in Table 12.

Table 12. The results of DIF analysis

Madde	χ2	df	p	UA
1	495.47	2	.00	0.0269
2	2033.71	2	.00	0.0307
3	74.08	2	.00	0.0121
4	6162.23	2	.00	0.1086
5	603.98	4	.00	0.0198
6	399.49	2	.00	0.0113
7	86.35	2	.00	0.0133
8	144.71	4	.00	0.0314
9	20.31	2	.00	0.0061
10	2211.70	2	.00	0.0218
11	649.22	2	.00	0.0166
12	710.93	2	.00	0.0191
13	**3.35**	**2**	**.19**	**0.0032**
14	236.89	2	.00	0.0071
15	110.76	2	.00	0.0180
16	182.21	4	.00	0.0380
17	363.73	2	.00	0.0152
18	63.39	2	.00	0.0147
19	1205.55	4	.00	0.0775
20	488.66	2	.00	0.0151

When Table 12 is examined, it is seen that the item parameters differ significantly between male and female students in all the other items in the mathematics subtest, except for the 13th item ($p < .05$).

Table 13 presents the DIF status of the 2012 HSEE Mathematics subtest items according to the effect size classification determined by Jodoin and Gierl.

Table 13. The items showing DIF and their effect size

The Group	The Items not Showing DIF	The Items Showing DIF		
		Negligible UA<.059	**Medium Level** .059 < UA < .088	**Significant level** UA >.088
Gender	13	1,2,3,5,6,7,8,9,10,11,12,14,15,16, 17,18,19,20	19	4

When Table 13 is examined, it is seen that item 19 shows DIF at a moderate level and item 4 shows DIF at a significant level in the mathematics subtest. It was determined that the 13th item did not show DIF (*p>.05*), and except for these three items, the other items in the test showed negligible DIF.

Within cognitive diagnostic models, uniform DIF is the case which poses consistently lower/higher correct answer probabilities for a group across all trait profiles. Non-uniform DIF, on the other hand, is that the probability of answering correctly is higher for a group in some latent profiles, while it is lower in some other attribute profiles in the same group (Li & Wang, 2015). The attribute profiles of male and female students regarding item 19 and their probability of answering correctly are given in Table 14.

Table 14. Success probabilities according to gender for item 19

Item 19

Bir yarışma programında, verilen her doğru cevaba +3 puan, her yanlış cevaba -2 puan verilmektedir.
Bu yarışmaya katılan Aysun, sorulan 5 sorunun tümünü cevaplamıştır. Yarışma sonunda 10 puan aldığına göre, Aysun kaç soruyu doğru cevaplamıştır?

A) 2 B) 3 C) 4 D) 5

The Group	Attributes Required for the Item	Attribute Pattern	Probabilities of Success
Female	α_1 ve α_3	A00	0.22
	α_1 ve α_3	A10	0.72
	α_1 ve α_3	A01	0.74
	α_1 ve α_3	A11	0.94
Male	α_1 ve α_3	A00	0.31
	α_1 ve α_3	A10	0.88
	α_1 ve α_3	A01	0.91
	α_1 ve α_3	A11	0.97

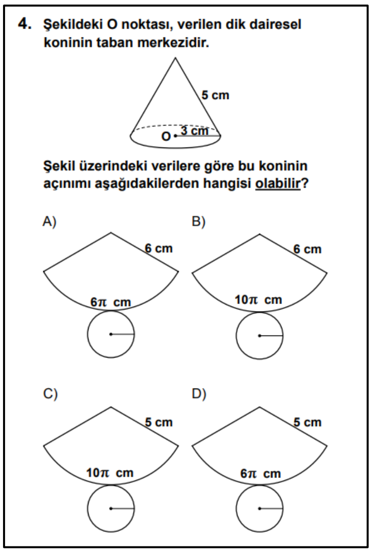
When Table 14 is examined, it is seen that the probability of answering the item correctly in item 19, which was determined to show moderate DIF, is higher for male students than for female students in all reduced attribute profiles. In this case, it can be stated that item 19 shows a uniform DIF in favor of male students.

As only one attribute is required for item 4, which was determined to show significant DIF, the success probability of students who do not have this attribute P(0) and those who have it P(1), can be thought of as the g and 1-s parameters in the DINA model, respectively (de la Torre, 2011). Hou et al. (2014) demonstrate that if the shift parameter (s) is small and the estimation parameter (g) is large for the focus group, the item shows uniform DIF, and as a result, they state that the probability of answering the item correctly in the focus group compared to the reference group, regardless of its implicit qualities is higher.

The attribute profiles and correct answer probabilities for item 4, which was determined to show significant DIF according to gender, are given in Table 15.

Table 15. Success probabilities according to gender for item 4

Item 4



4. Şekildeki O noktası, verilen dik dairesel koninin taban merkezidir.

Şekil üzerindeki verilere göre bu koninin açınımı aşağıdakilerden hangisi olabilir?

The Group	Attributes Required for the Item	Attribute Pattern	Probabilities of Success
Female	α_2	A0	0.64
	α_2	A1	0.94
Male	α_2	A0	0.51
	α_2	A1	0.91

In Table 15, when the success probabilities P(0) and P(1) are considered as the g and 1-s parameters in the DINA model respectively, it is seen that the estimation parameter (g) is higher and the shift parameter (s) is smaller for female students. Considering this, the results revealed that the probability of female students to answer the item correctly is higher than male students. In this case, it can be stated that item 4 shows a uniform DIF in favor of female students.

CONCLUSION, DISCUSSION AND SUGGESTIONS

This current study was carried out on the item answers of the 2012 HSEE 8th Grade Mathematics subtest. In the study, estimations of item parameters and student attribute profiles were obtained using the G-DINA model, and whether mathematics items showed DIF according to the gender variable within the scope of cognitive diagnosis models was examined.

General cognitive diagnostic models, such as the G-DINA model, control model fit at both test and item levels (Ravand, 2016). When the absolute and relative fit indices were examined in the current study, it was determined that the G-DINA model was compatible with the data at the test level. This indicates that there are “compensatory” relationships for some items and “non-compensatory” relationships for some other items among the attributes. General cognitive diagnostic models do not allow relationships between test-level attributes to be seen but allow different CDMs for different items within the same test. It is stated that the G-DINA model is particularly suitable in situations where the relationships between attributes may change depending on cognitive difficulties (Ravand, 2016). In the current study, the model fit was evaluated at the test level since the relationships between the "numbers", "geometry and measurement", "algebra", and "probability and statistics" attributes identified related to mathematics items were changeable and the cognitive structures of the items could not be understood, and the G-DINA model was observed to fit well at a high level. The G-DINA model, which is a saturated model, provides better model-data fit compared to other reduced models (de la Torre et al., 2015). However, reduced models are preferred under some conditions due to the principle of parsimony, which allows for more clear and understandable interpretations, requires small samples to obtain accurate estimations, and if it is not possible to distinguish between them, simpler models are preferred to complex models. Rojas et al., (2012) state that, unlike general models at the item level, reduced models such as DINA, DINO, ACDM, and NC-RUM are more interpretable in terms of the relationships between attributes. De la Torre (2011) suggests the Wald test as a statistical method to determine model fit at the item level, in other words, whether one of the reduced models can be used instead of the G-DINA model in the item. Accordingly, item-level model selection for items determined to require more than one attribute may be examined using the Wald test.

When the G-DINA model parameters are examined in the study, it is seen that the $\alpha_1 = [0000]$ attribute profile (61%) in which none of the four qualities defined is found in the student, and the $\alpha_{16} = [1111]$ attribute profile (17%) in which all the four attributes are present in the student (17%) are the most common attribute profiles. This finding is in line with other cognitive diagnosis studies (Lee & Sawaki, 2009b; Li, 2011; Ravand, 2016). In the study, it was determined that the tetrachoric correlation coefficients between the "Numbers", "Geometry and Measurement", "Algebra", and "Probability and Statistics" attributes varied between 0.69 and 0.99. High levels of positive correlations between attributes may be the reason for the backlog in the $\alpha_1 = [0000]$ and $\alpha_{16} = [1111]$ attribute profiles. Rupp et al. (2010) stated that the skewness in these attribute profiles may be due to the high positive correlation between attributes or the one-dimensionality of the measurements (Lee & Sawaki, 2009b). Here, what is meant by the one-dimensionality of the measurements is that if one of the required attributes is found, another attribute tends to be present or on the contrary, if one of the necessary attributes is lacking, the other attribute is also missing.

In the study, it was seen that 33% of the students have attributes related to the learning field of "probability and statistics". Accordingly, “probability and statistics” might be expressed as the easiest attribute. This quality was followed by "numbers" (28%), "algebra" (26%), and "geometry and measurement" (20%) relatively. Similarly, in the Turkish sample TIMSS 2011 study, it was found that 8th grade students found the questions in the fields of "algebra", "numbers", and "geometry" more difficult than the questions in the field of "data and probability" (Büyüköztürk et al., 2014). In his study, Atar (2011) applied descriptive and explanatory item response models to TIMSS 2007 Türkiye 8th grade mathematics data and created a linear logistic model with cognitive domain and subject domain variables to explain the differences in item difficulties. As a result of the analysis of this model, the cognitive domain and the subject area were found to influence item difficulty. When the subject area variable was examined, it was found that the items related to geometry were more difficult than the items related to "algebra" and "data and probability", and there was no statistically significant difference between the items related to "geometry" and the items related to numbers in terms of item difficulty. It may be misleading to consider these findings regarding the ease and difficulty of learning areas alone. Because while an item in the field of "geometry and measurement" requires routine algorithmic operations, another item in the field of "probability and statistics" may make estimation and comparison skills necessary. Hence, it is thought that the item difficulty level should be examined together with the cognitive skills, item structure and type, as well as the learning domain skills. In studies on cognitive diagnosis models, it is seen that the Q matrix is formed by combining the learning domain and cognitive domain skills (George & Robitzsch, 2014). However, although the alternative Q-matrix based on the combination of attributes in the two domains solves the methodological problem of not defining the model, this practice also changes the attributes used in the models. It is thought that the alternative Q-matrix creation method based on the combination of the two-level attributes may be preferred in cases where the cognitive areas and achievements that are desired to be measured with the items can be clearly defined. Additionally, instead of Bloom's taxonomy and the classifications that take it into account, the Math taxonomy (Smith et al., 1996) developed specifically for mathematics can be used in the classification of cognitive attributes related to mathematics. However, the fact that HSEE mathematics items are complex items to measure multiple acquisitions and that cognitive skills vary depending on the possible strategies used in the item make it difficult to consider the cognitive domain in cognitive diagnosis studies.

In the study, using the Wald test (de la Torre, 2011; Hou et al., 2014) within the scope of the G-DINA model, whether the 2012 HSEE Mathematics subtest items showed a gender-varying item function was examined. Accordingly, DIF analyses were carried out within the scope of CDM, in which latent attribute profiles were taken instead of total scores as matching criteria, and the 4th item showed a uniform DIF in favor of female students. On the other hand, it was found that item 19 showed a uniform DIF in favor of middle-level male students. In his study where Yıldırım (2015) examined whether the 2012 HSEE Mathematics subtest items showed DIF according to gender using MH and logistic regression (LR) methods, he determined that the 4th item showed moderate DIF in favor of female students and the 19th item in favor of male students in line with the current study. When the findings are compared with Yıldırım's (2015) study, it is seen that they overlap to a large extent, and the only difference is in the effect size classification of the 4th item. This difference is thought to be due to the use of different criteria (Zieky, 1993; Zumbo & Thomas, 1997; Jodoin & Gierl, 2001) for effect size classification. In addition, it may be suggested that the DIF determination approaches, in which the latent attribute profile is used as the matching criterion as a part of CDM, are similar to the traditional DIF determination approaches that use the total score as the matching criterion. A review of the literature reveals similar research findings (Zhang, 2006; Hou et al., 2014). Hou et al. in their study, in which they compared the DIF detection performance of the Wald test and traditional MH and SIBTEST methods, reported that the DIF detection performance of the Wald test was similar or superior to the MH and SIBTEST methods. In this study, it was stated that the Wald test is a promising approach in determining DIF within the scope of cognitive diagnostic models. Similarly, Zhang investigated the effectiveness of MH and SIBTEST methods together with two matching criteria (total score and attribute profile score) to determine DIF within the scope of CDM, and they were found to perform at the same level or better than the MH and SIBTEST methods (according to the total score approach) that utilize attribute profile score as a matching variable. However, as the use of different criteria for effect size classification (Zieky, 1993; Zumbo & Thomas, 1997; Jodoin & Gierl, 2001) leads to different DIF levels, it makes it difficult to decide on the compatibility between the methods.

Camilli and Shepard (1994) stated that there are two reasons for the emergence of DIF: item bias and item effect. Item effect indicates the actual differences between the groups on the characteristics to be measured with the item, while item bias indicates the differences unrelated to the measured construct resulting from the characteristics of the test items or the test conditions that are not fit for the purpose of the test (Zumbo, 1999). In his study examining the bias status of the 2012 HSEE Mathematics subtest items, Yıldırım (2015) investigated whether the DIF sources of the items 4 and 19 showing DIF that were found to have a strong consensus via the Delphi technique were related to the measured construct, and he found that the DIF in the 4th item was determined as item effect, and the DIF in item 19 as item bias. However, in the current study, which was carried out within the scope of cognitive diagnosis models, it was thought that the difference in the probability of correct answers to the 4th item, which was determined to show DIF according to gender as a result of DIF analysis, led to item bias while it was item effect in item 19.

In the Q-matrix, where the relationships between items and attributes were defined, a single attribute (geometry and measurement) was defined for item 4, and two attribute fields (numbers and algebra) were defined for item 19. Experts stated that the 4th item was prepared for the acquisition of "determines the basic elements of the cone, constructs it and draws the surface angles" in the geometric objects sub-learning area. They stated that item 19 was prepared for the acquisition of "solves linear equation systems with algebraic methods" in the field of algebra learning, but that the negative integers included in the item content required the attributes in the field of learning numbers. It was determined that the probability of answering the item correctly was quite high (58%) for the students who did not have the "geometry and measurement" field attribute, which was determined to be necessary for the 4th item in the Q-matrix. This indicates that other qualities or variables other than the "geometry and measurement" field attribute may have been effective for the solution of the item. Therefore, it may be assumed that DIF in item 4, which was determined to show a significant level of DIF, arises from a situation unrelated to the measured structure. In this case, the DIF in item 4 indicates item bias, not item effect resulting from real differences on the item and the attribute to be measured. When the success probabilities of item 19 depending on the latent attribute pattern were examined, it was determined that the success probability of the students who did not have the necessary attributes was relatively lower, while the "numbers" and "algebra" field attributes required for the item provided significant increases in the probability of answering correctly. When the success probabilities of male and female students according to the implicit attribute pattern were compared, there was a striking difference in the success probabilities. This indicates that the differentiation in item 19, which was determined to show a medium level of DIF, is due to the real differences in the attributes that are intended to be measured with the item. Therefore, the DIF in item 19 does not show the item bias arising from a situation unrelated to the construct measured by the item, but rather the item effect resulting from the actual differences in the item and the attribute to be measured. It is thought that with the DIF analyses considered within the scope of cognitive diagnosis models, the relationships between the sources of DIF and the attributes to be measured with the item can be defined more clearly, thus forming a statistical basis for item bias decisions. Moreover, DIF studies that will be carried out considering item attributes and multi-level data structure will provide richer information on DIF resources. Especially since the data in the field of education show multilevel data structure, it is thought that it would be interesting to consider multilevel DIF studies as a part of IRT within the scope of cognitive diagnosis models. In addition, it can be mentioned that there is an increasing interest in the use of machine learning methods in identifying potential DIF sources.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship and publication of this article.

Statements of publication ethics

We hereby declare that the study has not unethical issues and that research and publication ethics have been observed carefully.

Researchers' contribution rate

The study was conducted and reported with equal collaboration of the researchers.

Ethics Committee Approval Information

All stages of the study were carried out in accordance with ethical principles. Ready-made data were used within the scope of the study. The data were provided by the Ministry of National Education General Directorate of Measurement, Evaluation and Examination Services (Date: 09/11/2022-Number: E-57750415-622.03-63110119). The authors declare that they have no conflict of interest.

REFERENCES

Atar, B. (2011). Tanımlayıcı ve açıklayıcı madde tepki modellerinin TIMSS 2007 Türkiye matematik verisine uyarlanması. *Eğitim ve Bilim*, *36*(159).

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*(2), 179–197. https://doi.org/10.1007/BF02291262

Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı*. Pegem Akademi.

Büyüköztürk, Ş., Çakmak, K.E., Akgün, E.Ö., Karadeniz, Ş., & Demirel, F. (2014). *Bilimsel araştırma yöntemleri* (17. Baskı). Ankara: Pegem.

Büyüköztürk, Ş., Çakan, M., Tan, Ş., & Atar, H. Y. (2014). TIMSS 2011 ulusal matematik ve fen raporu–8. sınıflar. İşkur Matbaacılık, Ankara.

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

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265-289. https://doi.org/10.3102/10769986022003265

Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*, 19-38. https://doi.org/10.1111/j.1745-3984.2011.00158.x

de Ayala, R. J. (2009). *Theory and practice of item response theory*. Guilford Publications.

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

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115-130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*, 355-373. https://doi.org/10.1111/jedm.12022

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

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of Clinical Data From Cognitive Diagnosis Modeling Framework. *Measurement and Evaluation in Counseling and Development*, 1-16. https://doi.org/10.1177/0748175615569110

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics. Volume 26: Psychometrics* (pp. 979-1030). Amsterdam, The Netherlands: Elsevier. https://doi.org/10.1016/S0169-7161(06)26031-0

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495-515. https://doi.org/10.1007/BF02294487

Embretson, S.E. (1997). *Multicomponent response models*. In: *van der Linden,W.J., Hambleton, R.L. (Eds.),Handbook of Modern Item Response Theory.* New York: Springer, pp. 305–321. https://doi.org/10.1007/978-1-4757-2691-6_18

Fraenkel, J. R. & Wallen, N. E. *(2012). How to design and evaluate research in education* (Sekizinci Baskı). New York: McGraw-Hill.

George, A. C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling*, *56*(4), 405-432

Grønmo, L.S., Lindquist, M. ve Arora A. (2014). TIMMS Advanced 2015 Assessment Frameworks. *International Association for the Evaluation of Educational Achievement*. Mullis, I. V., & Martin, M. O.(Ed.).TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Haagenars, J., & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511499531

Hambleton, R. K., Swaminathan, H. ve Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage Publications.

Hartz, S., Roussos, L., & Stout, W. (2002). Skills diagnosis: Theory and practice. *User Manual for Arpeggio software. ETS*.

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

Junker, B.W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272. https://doi.org/10.1177/01466210122032064

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Lee, Y. W., & Sawaki, Y. (2009a). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, *6*(3), 172-189. https://doi.org/10.1080/15434300902985108

Lee, Y.-W., & Sawaki, Y. (2009b). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*, 239-263. https://doi.org/10.1080/15434300903079562

Lei, P. W., & Li, H. (2016). Performance of Fit Indices in Choosing Correct Cognitive Diagnostic Models and Q-Matrices. *Applied Psychological Measurement*, 1-13. https://doi.org/10.1177/0146621616647954

Leighton, J. P.& Gierl M. J. (2007). *Why Cognitive Diagnostic Assessment?* Leighton, J. P. Gierl M. J. (Ed). Cognitive Diagnostic Assessment for Education. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

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

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow***,** 9, 17-46.

Li, X., & Wang, W. C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, *52*(1), 28-54. https://doi.org/10.1111/jedm.12061

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020. https://doi.org/10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732. https://doi.org/10.1007/s11336-005-1295-9

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the Fit of Item Response Theory and Factor Analysis Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 333–356. https://doi.org/10.1080/10705511.2011.581993

Maydeu-Olivares. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*, 71–137. https://doi.org/10.1080/15366367.2013.831680

Maydeu-Olivares, A., & Joe, H. (2014). Assessing Approximate Fit in Categorical Data Analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*, 23-40. https://doi.org/10.1207/s15327906mbr3001_2

Milewski, G. B., & Baron, P. A. (2002). *Extending DIF methods to inform aggregate report on cognitive skills*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Ravand, H., & Robitzsch, A. (2015). Cognitive Diagnostic Modeling Using R.*Practical Assessment, Research & Evaluation*, *20*(11), 1-12.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, 1-18. https://doi.org/10.1177/0734282915623053

Robitzsch, A., Kiefer, T., George, A., C., & Ünlü, A. (2016). *CDM: Cognitive Diagnosis Modeling*. R Package Version 4.99-11. https://sites.google.com/site/alexanderrobitzsch/software.

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state of the art. *Measurement*, *6*(4), 219-262. https://doi.org/10.1080/15366360802490866

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

Smith, G., Wood, L., Coupland, M., Stephenson, B., Crawford, K., & Ball, G. (1996). Constructing mathematical examinations to assess a range of knowledge and skills. *International Journal of Mathematical Education in Science and Technology*, *27*(1), 65-77. https://doi.org/10.1080/0020739960270109

Ömür Sünbül, S. ve Kan, A (2015). Bilişsel tanı modellerinde parametre kestirimini ve sınıflama tutarlılığını etkileyen faktörlerin incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi.* 1-18. https://doi.org/10.16986/HUJE.2015014663

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational and Behavioral Statistics*, *10*, 55-73.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125-145. https://doi.org/10.1177/014662168400800201

Yıldırım, A. & Şimşek, H. (2011).*Sosyal bilimlerde nitel araştırma yöntemleri* (8. Baskı). Ankara: Seçkin Yayınları.

Yıldırım, H. (2015). *2012 yılı seviye belirleme sınavı matematik alt testinin madde yanlılığı açısından incelenmesi* [Yayınlanmamış yükseklisans tezi]. Gazi Üniversitesi.

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

Zieky, M. (1993) Practical questions in the use of DIF statistics in test development. In P.W. Holland ve H. Wainer (Ed.), *Differential item functioning* (s.337-347). Hillsdale NJ: Erlbaum.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF (Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science). *Canada University of British Columbia.*

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, National Defense Headquarters.*

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