ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 212-225

Impact of Retrofitting and Item Ordering on DIF

Lokman AKBAY *

Abstract

Richer diagnostic information about examinees' cognitive strength and weaknesses are obtained from cognitively diagnostic assessments (CDA) when a proper cognitive diagnosis model (CDM) is used for response data analysis. To do so, researchers state that a preset cognitive model specifying the underlying hypotheses about response data structure is needed. However, many real data CDM applications are adds-on to simulation studies and retrofitted to data obtained from non-CDAs. Such a procedure is referred to as retrofitting, and fitting CDMs to traditional test data is not uncommon. To deal with a major validity concern of item/test bias in CDAs, some recent DIF detection techniques compatible with various CDMs have been proposed. This study employs several DIF detection techniques developed based on CTT, IRT, and CDM frameworks and compares the results to understand the extent to which DIF flagging behavior of items is affected by retrofitting. A secondary purpose of this study is to gather evidence about test booklet effects (i.e., item ordering) on items' psychometric properties through DIF analyses. Results indicated severe DIF flagging prevalence differences for items across DIF detection techniques employing Wald test, Raju's area measures, and Mantel-Haenzsel statistics. The largest numbers of DIF cases were observed when the data were retrofitted to a CDM. The results further revealed that an item might be flagged as DIF in one booklet, whereas it might not be flagged in another.

Key Words: Differential item functioning, DINA model, retrofitting, booklet affect, cognitive diagnosis models.

INTRODUCTION

In educational practice, many large-scale tests focus on summative assessment, and their formative features are limited. Tests developed to diagnose examinees' strengths and weaknesses may provide rich information toward formative assessment and are referred to as cognitively diagnostic assessments (de la Torre & Minchen, 2014). To obtain diagnostic information, examinee responses obtained from such assessment procedures may be analyzed via statistical models known as cognitive diagnosis models (CDMs). Such diagnostic information may be considered as valuable feedback for students, teachers, and educational programs. Generally, CDMs are used to estimate examinees attribute-profiles that are defined by the mastery or nonmastery status of measured attributes. Rather than being just a coarse indicator of how examinees think about and complete educational tasks, CDM enables practitioners to identify and report finer grained attributes examinees use to complete such tasks.

As the test development procedure and response data hold the characteristics of cognitively diagnostic assessment (CDA), then, a successful CDM application providing detailed information to facilitate the explanation of examinee performance might be possible. In other words, a cognitive model specifying a structure of the data by means of theories or hypotheses is needed and must be set a priori (Gierl & Cui, 2008; Rupp & Templin, 2008). However, as reported by Gierl, Alves, and Majeau (2010), many CDM applications are adds-on to simulation studies and retrofitted to previous test data. Cognitive diagnosis retrofitting refers to the application of CDM as a psychometric model to response data from traditional testing procedures (Gierl & Cui, 2008).

More often than not, we come across the studies retrofitting traditional test responses to CDMs to determine examinee attribute-profiles. Examples of real data retrofitting studies include Choi, Lee, & Park (2015) and Terzi & Sen (2019). For a recent comprehensive review of the CDM applications, including retrofitting studies, readers may refer to Sessoms and Henson (2018). In conducting large-

* Asst. Prof., Istanbul University-Cerrahpasa, Istanbul-Turkey, lokmanakbay@istanbul.edu.tr, ORCID ID : 0000-0003-4026-5241

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

Received: 25.02.2021
Accepted: 28.03.2021

scale tests, it is aimed to reveal the cognitive ability levels of individuals in their study areas. One of the primary concerns in large-scale exams is the validity of assessment (Kane, 2013). The validity of a measurement tool is the degree to which it serves specified purposes and that it does not involve other features (Messick, 1995). Test bias is one of the severe factors threatening the validity of a test. Bias is observed when examinees' test scores in different subgroups contain group-dependent systematic errors (Camilli & Shepard, 1994). Differential item functioning (DIF) detection is a useful tool for identifying item bias. DIF is defined as the differentiation of the probability of answering an item correctly among individuals who are in different subgroups but have the same ability level (Zumbo, 2007). In other words, DIF arises when an item's response function differs from one group to another.

When an item is diagnosed by a specific DIF technique, content domain and measurement experts examine the items to understand whether the item offers a systematic advantage in favor of any subgroup. This systematic advantage is referred to as item bias, and DIF analysis is a crucial step in item bias examination. Various statistical DIF detection techniques based on classical test theory (CTT) and item response theory (IRT) are used to identify DIF items. These techniques include Mantel-Haenszel (Holland & Thayer, 1988), Logistic Regression (Swaminathan & Rogers, 1990), IRTLR tests (Thissen & Steinberg, 1988), Lord's $\chi 2$ test (1980), and the MIMIC model (Jöreskog & Goldberger, 1975; Woods, 2009). Recently, DIF detection techniques for cognitive diagnosis modeling framework have also been proposed (Hou, Terzi & de la Torre, 2020; Ma, Terzi & de la Torre, 2021). For example, Hou, de la Torre, and Nandakumar (2014) proposed a DIF detection method based on the Wald test that is compatible with the deterministic inputs, noisy "and" gate (DINA: Junker & Sijtsma, 2001) model. In this study, DIF detection techniques developed based on CTT, IRT, and CDM frameworks are employed. Namely, Mantel-Haenszel (Holland & Thayer, 1988), Raju's (signed) area measures (1988, 1990) and Wald test for DIF (Hou, de la Torre & Nandakumar, 2014) are employed.

In light of the above discussion, the primary purpose of this study is to examine the psychometric properties of a test through DIF analyses. Specifically, DIF flagging patterns of three DIF detection techniques, namely Mantel-Haenszel, Raju's area measures, and Wald test for DIF, are examined in terms of pattern consistency/similarity when the cognitive model specifying the data structure and psychometric model directing the psychometric analysis are different. In other words, DIF flagging patterns of the three DIF detection techniques were examined when response data are retrofitted. For this purpose, real data from a large-scale assessment are used. The data were collected using two booklets (i.e., Booklets A and B), and the subgroups of DIF analyses were based on variables gender and booklet type.

Another important issue on large-scale testing is the use of different booklets in test administration. Regarding the effect of using different types of booklets on the examinee achievement, testing agencies such as Measurement, Selection, and Placement Center (ÖSYM) argue that random assignment of test items to the booklets does not have any impact on examinees' achievement (2011). On the contrary, some experts claim that the positions of the items in the booklet could affect examinee performance by affecting anxiety and motivation levels, from which the estimates of test's psychometric properties may be affected (Middle East Technical University-METU, 2011; Ankara University, 2011). Although revealing the effect of the booklet on a single examinee is not feasible, the booklet effect on estimates of tests' psychometric properties can be statistically examined. Then, the secondary purpose of this study is to examine impact of the booklet on DIF analyses. Specifically, gender DIF flagging pattern of items across Booklets A and B is documented. Therefore, both the booklet effects and impact of retrofitting on real testing situations are examined, and the compatibility of Wald test based DIF detection under DINA model with more traditional DIF detection techniques is emphasized.

Purpose of the Study

Below research problems are addressed in this study:

- Do the DIF detection techniques developed based on CTT, IRT, and CDM frameworks yield compatible results (focusing on the cases where data are retrofitted)?

- Do the DIF flagging items differ across test booklets with different item ordering? In other words, do DIF analysis results get affected by the order of test items?

Dif Detection Techniques

Mantel-Haenszel technique for DIF detection

This CTT based DIF detection technique was proposed by Holland and Thayer (1988) using the statistic developed by Mantel and Haenzsel (1959). This technique is referred to as Mantel-Haenzsel DIF technique and examines whether item responses are independent of group membership after conditioning on the observed total score. The test statistic in this technique asymptotically follows a chi-square (χ^2) distribution with 1 degrees of freedom so that the statistic is compared against the chi-square distribution. To obtain the test statistic (χ^2_{MH}), for all total scores from 1 to $J-1$, N_m examinees are classified into 2×2 contingency tables, where J is the total number of items in the test and N_m is the number of examinees obtained a total score of m.

Table 1. A 2×2 Contingency Table Conditioned on the Total Score of m

Correct response to item *j*	Incorrect response to item *j*	Total response to item *j*
C_{Fm}	I_{Fm}	N_{Fm}
C_{Rm}	I_{Rm}	N_{Rm}
$N_{Cm} = C_{Fm} + C_{Rm}$	$N_{Im} = I_{Fm} + I_{Rm}$	$N_m = N_{Fm} + N_{Rm} = N_{Cm} + N_{Im}$

Note. C_{Fm} is the number of examinees who correctly responded to item *j* in the focal group; I_{Fm} is the number of examinees incorrectly responded to item *j* in the focal group; N_{Fm} is the total number of examinees with a total score of *m* in the focal group; C_{Rm} is the number of examinees correctly responded to item *j* in the reference group; I_{Rm} is the number of examinees incorrectly responded to item *j* in the reference group; N_{Rm} is the total number of examinees with a total score of *m* in the reference group; N_{Cm} is the total number of examinees with a total score of *m* who correctly responded to item *j*; N_{Im} is the total number of examinees with a total score of *m* who incorrectly responded to item *j*; and N_m is the total number of examinees with a total score of *m*.

Based on the information obtained from 2×2 contingency tables, the below formula is used to obtain test statistic:

$$\chi^2_{MH} = \frac{\left\{\left|\sum_{m=1}^{J-1}[C_{Rm} - E(C_{Rm})]\right| - 0.5\right\}^2}{\sum_{m=1}^{J-1} Var(C_{Rm})}, \tag{1}$$

where

$$E(C_{Rm}) = \frac{N_{Rm}N_{Cm}}{N_m} \tag{2}$$

and

$$Var(C_{Rm}) = \frac{N_{Rm}N_{Fm}N_{Cm}N_{Im}}{N_m^2(N_m - 1)}. \tag{3}$$

Raju's (Signed) area measures for DIF detection

This DIF detection technique is based on item response curves (IRCs) defined by the item parameters obtained under one- two-, or three parameter logistic models. For a dichotomously scored item, unidimensional three-parameter logistic model is defined as

$$P_j(\theta) = \gamma_j + (1 - \gamma_j)\left[1 + \exp\{-1.7\alpha_j(\theta - \beta_j)\}\right]^{-1}, \tag{4}$$

where $P_j(\theta)$ is the probability of correctly answering item *j* when examinee's continuous ability level is θ; γ_j is the pseudo-guessing parameter of item *j*; α_j is the discrimination parameter of item *j*; θ is the continuous ability level; and β_j is the difficulty parameter of item j. Two- parameter logistic model can be derived from the above function by setting γ_j to zero. Similarly, one-parameter logistic model is derived by setting γ_j to zero and α_j to an estimated constant. This estimated discrimination parameter is fixed for all items in the test.

For one- two-, or three-parameter logistic models, Raju's (signed) area measure is the area between the IRCs defined by the estimated item parameters of focal and reference groups (Raju, 1988, 1990). As stated by Raju (1988, 1990) when the pseudo-guessing parameters of the IRF of subgroups for three-parameter logistic models are not equal, the area between the two item characteristic curves becomes infinite. Therefore, to avoid this problem, he suggests constraining the lower asymptote (i.e., pseudo-guessing parameter) to a fixed value. Based on this technique, DIF is examined by comparing the computed area between the item response curves to the determined critical values.

Given the item response functions of focal and reference groups,

$$F_F(\theta) = \gamma_{Fj} + (1 - \gamma_{Fj})\left[1 + \exp\{-1.7\alpha_{Fj}(\theta - \beta_{Fj})\}\right]^{-1} \tag{5}$$

and

$$F_R(\theta) = \gamma_{Rj} + (1 - \gamma_{Rj})\left[1 + \exp\{-1.7\alpha_{Rj}(\theta - \beta_{Rj})\}\right]^{-1}, \tag{6}$$

the area between the curves determined by the functions is calculated by taking the integral of the absolute differences

$$Area = \int_{-\infty}^{\infty} |(F_R - F_F)| d\theta. \tag{7}$$

Then, based on the null hypothesis that the true area is zero, a test statistic Z corresponding to the measured area is computed and compared against standard normal distribution. Readers may refer to Raju (1990) for details on the computation of the Z statistics.

Wald test for DIF detection under DINA model

One of the most parsimonious CDMs is the DINA model (Junker & Sijtsma, 2001), which is used to predict the probability of correctly answering an item as a function of individuals' discrete attributes' mastery status and item parameters (Li, 2008). Based on the DINA model, examinees' attribute profiles indicating mastered and nonmastered attributes are estimated. Regardless of the number of attributes measured by the test and the number of attributes required by an individual item, for DINA model, two item parameters are estimated. These parameters are referred to as guessing and slip parameters (de la Torre, 2009). Guessing parameter of item j (g_j) is the probability of successful response of an examinee who has not mastered at least one of the attributes that are required to correctly answer item j. Likewise, the slip parameter of item j (s_j) is the probability of incorrectly responding to item j when an examinee has already mastered all required attributes required by the item (de Carlo, 2012; de la Torre, 2009). These two parameters are mathematically defined as

$$g_j = p\left[X_{ij} = 1 | \eta_{ij} = 0\right] \tag{8}$$

and

$$s_j = p\left[X_{ij} = 0 | \eta_{ij} = 1\right], \tag{9}$$

where g_j is guessing parameter of item j; s_j: slip parameter of item j; η_{ij} is ideal response (i.e., when $s_j = g_j = 0$) of examinee i to item j.

Given the item parameters, the DINA model item response function (i.e., probability of correctly responding to given item) is defined as

$$P(X_{ij} = 1|\boldsymbol{\alpha}_l) = g_j^{(1-\eta_{jl})}(1-s_j)^{\eta_{jl}}, \tag{10}$$

where X_{ij} is the observed response of examinee i to item j; $\boldsymbol{\alpha}_l$ is attribute vector l among 2^K attribute vectors formed by K measured attributes; η_{il} is the ideal response of an examinee when his/her attribute vector is $\boldsymbol{\alpha}_l$.

First of all, in CDM context, DIF refers to the difference in the success probability of reference and focal groups with the same attribute mastery patterns (Hou et al., 2014). Under the DINA model, DIF is observed for item j when $\Delta g_j = g_{Fj} - g_{Rj} \neq 0$ and/or $\Delta s_j = s_{Fj} - s_{Rj} \neq 0$, where F and R stand for focal and reference groups, respectively. When Δg_j and Δs_j have the same sign, the DIF referred to as uniform; otherwise, it is called non-uniform DIF. Wald test DIF for the DINA model tests the significance of the joint differences between the item parameters of the subgroups:

$$W_d = (C\hat{v}_j)'(C\hat{\Sigma}_j C')^{-1}(C\hat{v}_j), \tag{11}$$

where $\hat{v}_j$ is an item parameter column vector of $(g_{Fj}, s_{Fj}, g_{Rj}, s_{Rj})^T$; $\hat{\Sigma}_j$ is asymptotic variance-covariance matrix associated with the subgroups' item parameter estimates; and C is the contrast matrix of $\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$. In this test, W_d asymptotically follow a chi-square (χ^2) distribution with 2 degrees of freedom, and the tested null hypothesis is $C\hat{v}_j = 0$.

METHOD

Sample

The data used in this study were obtained from a 19-item mathematic section of the high school admission exam (TEOG). More specifically, the data are the responses of high school applicants who took the test in 2013 in Ankara, Turkey. It should be noted here that rather than answering any specific research questions raised about this specific exam, this study employed this data set to mimic real life conditions where the data analysis may or may not flag DIF items. In other words, this dataset is used in this simulation-like study rather than using simulated data that may not truly reflect real life conditions. For the current study, 100 datasets were randomly drawn from the entire data, which consist of 39,146 male and 37,318 female examinees' responses to 19 multiple-choice mathematics items. The sample size for each data was fixed to 1,000 in order to obtain stable item parameter estimates under the DINA and IRT models for both focal and reference groups. This sample size is sufficient for unbiased and accurate estimation of the DINA model parameters (see De la Torre, Hong, & Deng, 2010) as well as unidimensional three-parameter logistic (3PL) model parameters (de Ayala, 2009, p. 130). In the study, Ox-Edit program (Doornik, 2003) was used for random sample drawings, and DIF analyses were conducted via R-programming (R Core Team, 2016).

Table 2: Descriptive Statistics by the Booklet Type

	Booklet A			Booklet B		
	Male	*Female*	*Total*	*Male*	*Female*	*Total*
Number of examinees	20,076	18,869	38,945	19,070	18,549	37,619
Number of items	19	19	19	19	19	19
Mean	8.49	9.499	8.979	8.801	9.776	9.28
Variance	26.099	25.471	26.048	24.854	23.742	24.558
Standard deviation	5.108	5.047	5.104	4.988	4.873	4.955
Skewness	-0.694	-0.908	-0.894	-0.755	-0.97	-0.893
Kurtosis	0.552	0.288	0.417	0.599	0.35	0.447

As stated above, because this study has no specific interest in examining either test items in detail or examinee achievement, descriptive statistics are not thoroughly discussed. Rather, descriptive statistics for each gender group for the A and B test booklets are summarized in Table 2.

Dimensionality

To be able to apply Raju's area statistic based on the unidimensional IRT model, the data need to be unidimensional. So, dimensionality was checked through exploratory factor analysis conducted via SPSS, and the results confirmed the unidimensionality. The results of this analysis are presented in Table 3.

Table 3. Findings of Exploratory Factor Analysis

	1st Dimension	2nd Dimension	3rd Dimension	4th Dimension	5th Dimension
Explained variance	.33	.06	.05	.04	.04
Cumulative explained variance	.33	.39	.44	.48	.52

Model Selection

To be able to retrofit the data to a CDM, an item-attribute specification matrix, namely, Q-matrix was developed after establishing the attributes measured by the test. The attributes were set, and the Q-matrix was constructed by mathematics education experts. The model fits statistics indicated an acceptable fit of the data to the DINA model so that Wald test based DIF detection under the DINA model was conducted. In terms of unidimensional models, data were fitted to the Rasch, 1PL, 2PL, and 3PL IRT models for model selection. It should be recalled that the only difference between the Rasch model and 1PL model is the common item discrimination index. In particular, item discrimination is fixed to 1.00 for all items under the Rasch model. On the contrary, under the 1PL model, a common discrimination parameter is estimated from the data and fixed across all items in the test. Model selection yielded that the 3PL model best fitted to the data, and the model selection results were presented in the results section.

Analysis

In order to facilitate the analyses and interpretation of the analyses, the order of the items in different booklets was rearranged before conducting the analyses for which booklet A was taken as reference. Each of the 100 datasets was obtained from the entire examinee response data, and these data sets were analyzed through the Wald test, Raju's area measures, and Mantel-Haenzsel DIF detection techniques for gender groups. To understand the impact of booklet type on estimated item parameters (i.e., the impact of item ordering on psychometric properties of a test), DIF analyses were conducted on booklet A and B separately, and the results were compared. To perform the analyses, Ox-Edit program for the Wald test cases and the difR package (version 4.6) developed by Magis, Beland, and Raiche (2015) were used for Raju's area measures and Mantel-Haenzsel DIF detection cases. Comparing the obtained test statistics to corresponding relevant statistical distributions, *p*-values were computed and reported to compare and contrast DIF detection results of different techniques and their variation by test booklets. Therefore, by comparing and contrasting the obtained p-values to the significance levels of $\alpha = .01$ and $\alpha = .05$, DIF flagging rates across two booklets and different DIF detection techniques were examined.

RESULTS

To determine which IRT model to employ for the Raju's area measure DIF technique, a model selection analysis was conducted to select one from one-, two-, and three-parameter logistic models. Because all four models are nested, a deviance test (i.e., likelihood ratio test) test is also conducted along with consideration of Akaike's information criterion (AIC) and Bayesian information criterion (BIC) for model selection. The test statistics and the test results are given in Table 4, which indicate that 3PL model is the best fitting model among all four. As discussed by Raju (1988, 1990), area measures for DIF detection are computed after fixing the lower asymptote. For this study, because all items in the test were multiple-choice with four options, theoretically constraining the pseudo-guessing parameter to 0.25 was meaningful. Accordingly, for the purpose of employing Raju's area measures DIF detection technique, 3PL model pseudo-guessing parameters were set to 0.25 across all items.

Table 4. Data-Model Fit Statistics

Model	AIC	BIC	Loglikelihood	-2*x*Loglikelihood	*df*
Rasch	820747.5	820910.3	-410354.7	-----------	---
1PL	811908.6	812080.0	-405934.3	8840.85*	1
2PL	796224.6	796550.3	-398074.3	15719.99*	18
3PL	788745.1	789233.6	-394315.5	7517.56*	19

Note: * p<.001, AIC is Akaike information criterion; BIC is information criterion; and *df* stands for degrees of freedom.

One of the main aims of this study was to examine the variation in DIF-flagging prevalence of the test items when analyzed under different psychometric models. This study especially focused on the variation in DIF analysis results when the data were retrofitted to a CDM such as DINA model. Thus, DIF flagging rates of three DIF techniques employed for CTT, IRT, and CDM-based psychometric models were examined, and the results at α = .05 and α = .01 levels were summarized in Table 5 and 6, respectively. For example, at α-level of .05, item-1 was flagged as DIF-item by Raju's area measures 22 out of 100 times in booklet A and 32 out of 100 times in booklet B conditions. Likewise, the number of times this item was flagged as DIF-item at α-level of .01 were 5 and 14 under booklet A and B, respectively.

Table 5. Null Hypotheses Rejection Rates of the DIF Detection Techniques at α = .05

	Psychometric models used as a basis for DIF analyses					
	Wald test for DINA		*Raju's area for 3PL*		*Mantel-Haenzsel for CTT*	
Items	Booklet A	Booklet B	Booklet A	Booklet B	Booklet A	Booklet B
Item 1	.51	.79	.22	.32	.12	.40
Item 2	.16	.25	.23	.32	.05	.05
Item 3	.20	.23	.55	.64	.63	.69
Item 4	.39	.39	.41	.27	.48	.47
Item 5	.10	.15	.02	.06	.16	.30
Item 6	.91	.79	.26	.29	.42	.35
Item 7	.65	.62	.17	.20	.06	.02
Item 8	.49	.27	.00	.00	.05	.01
Item 9	.31	.56	.07	.04	.33	.44
Item 10	.62	.38	.11	.14	.17	.13
Item 11	.38	.12	.24	.21	.05	.05
Item 12	.53	.66	.45	.25	.86	.82
Item 13	.34	.55	.17	.15	.07	.25
Item 14	.92	.92	.09	.11	.66	.61
Item 15	.19	.17	.12	.37	.06	.07
Item 16	.48	.35	.46	.39	.06	.04
Item 17	.96	.81	.49	.64	.70	.44
Item 18	.66	.76	.37	.43	.53	.65
Item 19	.07	.09	.68	.77	.26	.24

The rejection rates of the null hypotheses given in Tables 5 and 6 were obtained by comparing the observed p-values of the analyses to the critical values of .05 and .01, respectively. Thus, it is not clear whether the null hypotheses were rejected with a p-value of .051 or .999. Therefore, in addition to the null hypotheses rejection rates presented in the abovementioned tables, boxplots were also created based on the p-values obtained from analyses of 100 data sets for each of the booklets. These boxplots are presented in Figure 1, in which horizontal lines indicate the null hypothesis rejection levels of .01 and .05.

By looking at the tables, severe differences in the prevalence of DIF flagging for an item can be observed across different DIF techniques. First of all, numbers of DIF cases are the largest for Wald test DIF detection under the DINA model with grand mean ratios of .47 and .31 when $\alpha = .05$ and $\alpha = .01$, respectively. Although they are not quite different from the Mantel-Haenzsel results, the smallest grand means for DIF flagging rates (mean rates of .28 and .11 when $\alpha = .05$ and $\alpha = .01$, respectively) are observed for Raju's area measures under 3PL model. Lastly, the Mantel-Haenzsel DIF technique yielded a grand mean null hypotheses rejection rates of .31 and .16 under $\alpha = .05$ and $\alpha = .01$, respectively.

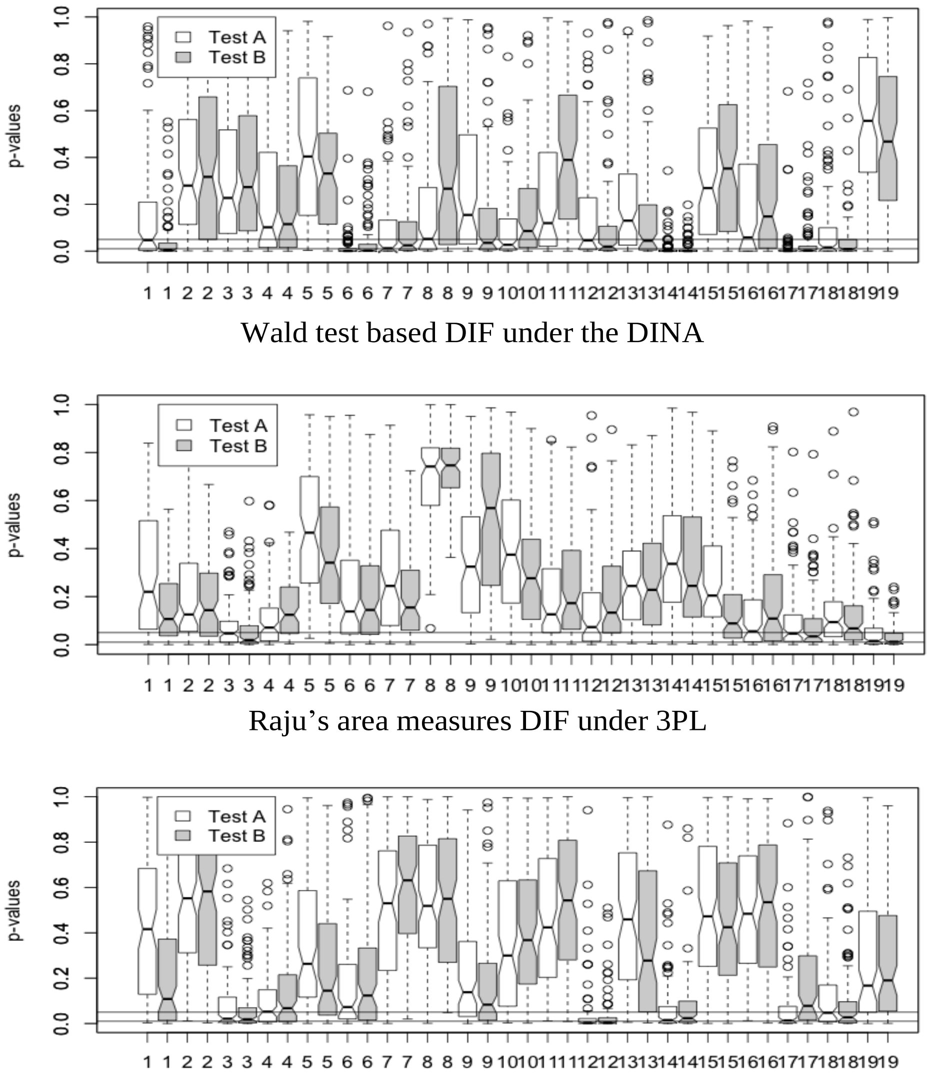
In terms of pairwise comparisons of DIF techniques, the largest differences in the DIF flagging ratios were observed between the Wald test and Raju's area measures. Relatively large differences in the prevalence of DIF flagging are observed for 13 out of 19 items (items 1, 3, 6, 7, 8, 9, 10, 12, 13, 14, 17, 18, and 19). For this comparison, the largest difference was observed for items 14A and 14B with differences of $.92 - .09 = .83$ and $.81 - .02 = .79$ for $\alpha = .05$ and $\alpha = .01$ cases, respectively. Further, in comparison of the DIF flagging ratios for the Wald test and Mantel-Haenzsel techniques, large differences were observed for 11 items (items 1, 3, 6, 7, 8, 10, 12, 13, 14, 16, and 17). In this comparison, the largest differences in ratios were observed for items 7B and 6A with differences of $.62 - .02 = .60$ and $.74 - .17 = .57$ when $\alpha = .05$ and $\alpha = .01$, respectively. When comparing the rejection rates of Raju's area measures and Mantel-Haenzsel techniques, the gaps between the ratios were relatively smaller. Nevertheless, five items (items 9, 12, 14, 16, and 19) were reported to have large differences in terms of the ratio of being flagged as DIF items. In this case, the largest ratio differences were reported for item 12B with a difference of $.82 - .25 = .57$ and $.63 - .09 = .54$ for $\alpha = .05$ and $\alpha = .01$ conditions, respectively.

Table 6. Null Hypotheses Rejection Rates of the DIF Detection Techniques at $\alpha = .01$

	Psychometric models used as a basis for DIF analyses					
	Wald test for DINA		*Raju's area for 3PL*		*Mantel-Haenzsel for CTT*	
Items	Booklet A	Booklet B	Booklet A	Booklet B	Booklet A	Booklet B
Item 1	.32	.62	.05	.14	.03	.19
Item 2	.01	.09	.08	.07	.00	.01
Item 3	.05	.09	.26	.36	.37	.42
Item 4	.21	.19	.18	.05	.32	.28
Item 5	.01	.05	.00	.01	.06	.14
Item 6	.74	.68	.09	.12	.17	.19
Item 7	.48	.33	.06	.05	.01	.00
Item 8	.31	.17	.00	.00	.01	.00
Item 9	.16	.34	.01	.00	.10	.21
Item 10	.33	.20	.02	.03	.06	.04
Item 11	.19	.07	.07	.06	.00	.01
Item 12	.31	.40	.19	.09	.58	.63
Item 13	.16	.37	.03	.03	.01	.07
Item 14	.78	.81	.04	.02	.41	.34
Item 15	.11	.08	.07	.16	.02	.03
Item 16	.44	.23	.24	.21	.00	.01
Item 17	.79	.66	.20	.21	.47	.15
Item 18	.46	.54	.14	.15	.31	.37
Item 19	.04	.04	.39	.48	.08	.09

The secondary purpose of this study was to investigate the booklet effect, if any, on estimated item parameters via DIF detection techniques. Because the DIF is examined through variations of items' psychometric properties, variation in observed DIF results across test booklets may be considered as empirical evidence to argue that item order in a test affects items' estimated parameters. When the Wald test DIF results for the DINA cases were examined, clear variations in DIF flagging rates of this technique for two test booklet conditions were observed. Specifically, when $\alpha = .05$ was considered, DIF flagging rates of seven items (items 1, 8, 9, 10, 11, 13, and 16) were substantially different. Even though the significance level was reduced to α = .01, five out of these seven items (items 1, 8, 9, 13, and 16) were flagged as DIF-items with notably different flagging rates. Similarly, Raju's area measures DIF flagging rates of four items (4, 12, 15, and 17) were relatively different across two test booklet conditions. Even under a more conservative α-level (i.e., α = .01), items four and 12 were still slightly diversified. Lastly, when detecting DIF items via the Mantel-Haenszel technique, the difference in DIF flagging rates of four items (items 1, 5, 13, and 17) came to the forefront. Among these four, items 1 and 17 remained diversified in terms of being flagged as DIF items under the α-level of .01.

Furthermore, Figure 1 was also used to explore the relationships between the booklets with respect to DIF flagging behavior. Boxplots in this Figure were plotted with notches, where lack of overlap between the notches of the boxplots for booklets A and B indicates that the median scores specified in these box plots are different (Chambers, Cleveland Kleiner, & Tukey, 1983). These plots in Figure 1 yielded compatible results from those presented in Tables 5 and 6. Specifically, the notches of the boxplots for booklets A and B did not have any overlap for items 1, 8, 9, 10, 11, and 13 when the DIF detection technique was the Wald test DIF for DINA. Similarly, when Raju's area measure and Mantel-Haenszel DIF detection techniques were employed, boxplot notches did not overlap for items 1, 4, 9, and 15; and items 1, 5, 13, and 17, respectively. Based on the above results, it is evident that booklet type yielded different outcomes from DIF analyses.



Wald test based DIF under the DINA

Raju's area measures DIF under 3PL

Mantel-Haenszel DIF under CTT

Figure 1. Boxplots of the p-values computed for DIF hypothesis testing.

DISCUSSION and CONCLUSION

In practice, many large-scale tests focus on summative assessments providing coarse test scores that provide limited formative information. Analyzing the data collected from cognitively diagnostic assessments (CDA) by CDMs may offer richer diagnostic information about examinees' cognitive strengths and weaknesses. Specifically, CDM enables practitioners to identify and report finer grained attributes examinees use to complete cognitive tasks. However, Gierl and Cui (2008) and Rupp and Templin (2008) state that a cognitive model specifying theories or hypotheses related to the structure of the data must be set. Yet, many real data CDM applications are adds-on to simulation studies and *retrofitted* to data already collected (Gierl, Alves, & Majeau, 2010; Terzi & Sen, 2019). Therefore, more often than not, practitioners fit CDMs to traditional test responses.

A major validity concern arises in large-scale assessments when item/test bias occurs, and DIF detection is a useful method for dealing with this validity thread. Various statistical techniques based on CTT and IRT are used to identify DIF-items. Up to date, DIF detection techniques that are compatible with CDMs, such as Wald test DINA DIF detection technique (Hou, de la Torre, & Nandakumar, 2014; Hou, Terzi, & de la Torre, 2020), have been proposed. In this study, DIF detection techniques developed based on CTT, IRT, and CDM frameworks are employed, and the results are compared to derive conclusions about the compatibility of the results. It is particularly important to understand how tests' psychometric properties are affected in retrofitting. Therefore, this study aimed to examine the psychometric properties of a test through DIF analyses. For this purpose, real data from a large-scale assessment were used. Because the dataset was collected via two test booklets with different item ordering, this study also examined the booklet impact on estimated item parameters through DIF analyses across gender groups were conducted on booklet A and B.

Results indicated severe DIF flagging prevalence differences for items across different DIF techniques. The largest numbers of DIF cases were observed under the DINA retrofitting, whereas comparably less frequent DIF cases observed when Raju's area measures under 3PL model and Mantel-Haenzsel DIF detection technique based on CTT were employed. One of the presumptive reasons for this result is that the original exam was not developed for CDA purposes. Specification of attributes to be measured by the test, development of items assessing the attribute set, and construction of the Q-matrix to establish a precise relationship between items and attributes are the key points for obtaining accurate information from a test in the CDA framework. Thus, the alignment of items and attributes in a test is a crucial step for enhancing the benefit of diagnostic assessment. In many cases, not specific for the test and data used in this study, psychometric properties of a test may not be accurately determined when data are collected via an achievement test that was not developed based on CDA.

Further results were obtained with respect to the booklet effect on items' psychometric properties through DIF detection techniques. When the Wald test DIF results for the DINA were examined, clear variations in DIF flagging rates of this technique for the two test booklet conditions were observed. Although the alterations of DIF analysis results across two booklets were not as high, DIF flagging rates of Raju's area measures and Mantel-Haenzsel techniques resulted in a similar pattern. Thus, it may be concluded that different booklets have an impact on the estimated psychometric properties of items such that these differences produce variant DIF patterns on a test. In the literature, there are studies suggesting that changes in item positions change the difficulty level of the items (Kingston & Dorans, 1984). In addition, it is also known that the speed responding to an item, fatigue, and exam experience can also lead to DIF. Thus, variations in items response speed, strategies used for response generation, cognitive effort exertion rate, and fatigue across subgroups may yield variation in estimated item parameters as item order changes in a test. Therefore, as the differences in the estimated item parameters for the subgroups increases due to the sequence of items in a test, items may be flagged by DIF detection techniques. Therefore, even if item ordering changes across booklets, these changes in item locations should not be dramatic to minimize item order effect on DIF and eventually on test scores.

REFERENCES

Ankara University (2011). *Ankara Üniversitesi Eğitim Bilimleri Fakültesi'nin YGS Hakkında Görüşü.* Retrieved November 30, 2015, form https://dahilibellek.wordpress.com/2011/04/12/ankara-ebf-ygs/

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

Chambers, J.M., Cleveland, W.S., Tukey, P.A., Kleiner, B. (1983). *Graphical Methods for Data Analysis.* Wadsworth International Group, the University of Michigan.

Choi, K. M., Lee, Y.-S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science, & Technology Education, 11,* 1563–1577.

de Ayala, *R. J. (2009). The theory and practice of item response theory.* The Guilford Press, New York, NY.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 47,* 115-127

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement 47,* 227–249.

de la Torre, J., & Minchen N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis framework. *Psicologia Educative 20,* 89-97

De Carlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36,* 447-468.

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, 10,* 318-341. doi:10.1080/15305058.2010.509554

Gierl, M.J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement, 6,* 263-275.

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

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.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70,* 631-639.

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.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8(2),* 147-154.

Li, F. (2008) *Modified higher-order DINA model for de11tecting differential item functioning and differential attribute functioning.* Unpublished doctoral dissertation University of Georgia, USA.

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

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.

Magis, D., Beland, S., & Raiche, G., (2015). *difR: Collection of methods to detect dichotomous differential item functioning (DIF).* R package version 4.6.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Middle East Technical University. *(2011). 2011 Yılı Yükseköğretime Geçiş Sınavı Hakkında ODTÜ Eğitim Fakültesi Görüşü.* Retrieved November 30, 2015, form *http://fedu.metu.edu.tr/sites/fedu.metu.edu.tr/files/ygs2011hkegitimfakultesigorusu_28_4_2011_v2.pdf*

Measurement, Selection, and Placement Center. *(2011). Adaya özgü soru kitapçığı.* Retrieved December 29, 2015, file://localhost/from http/::www.osym.gov.tr:belge:1-12431:adaya-ozgu-soru-kitapcigi-21032011.html

R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. *URL https://www.R-project.org/.*

Raju, N. S. (1988). The area between two item characterıstıc curves. *Psychometrika, 3,* 495-502.

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

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.

Sessoms, J. & Henson, R. A. (2018). Applications of Diagnostic Classification Models: A literatüre review and critical commentary. *Measurement: Interdisciplinary research and persperctives, 1,* 1-17.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4),* 361-370.

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),* 1-11.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, 99(1), 118.

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement, 32(7),* 511-526.

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.

Veriye Sonradan Model Eklemenin ve Madde Sıralamasının DMF Üzerindeki Etkileri

Giriş

Çoğu geniş ölçekli testler özetleyici değerlendirmeye yönelik olup genel ve özet puanlarla ölçülen özelliğin testi alanlardaki seviyesini ortaya koymakta ve biçimlendirici değerlendirme çerçevesinde oldukça sınırlı bilgi sağlayabilmektedir. Bilişsel tanılama yapabilmek adına geliştirilen testlerin sonuçları bilişsel tanı modelleri (BTM) aracılığıyla analiz edildiğinde ise testi alanların bilişsel niteliklere sahip olma ya da olmama durumları ile ilgili zengin tanısal geri dönütler elde edilebilir. BTM ile yapılan analizler, testi alanların test içerisinde sunulan bilişsel görevleri tamamlamak için kullandıkları küçük boyutlu ve ayrıntılı bilişsel niteliklerin tanımlamasını ve testi alanlarda bulunup bulunmama durumlarının belirlenmesini sağlar. Gierl ve Cui (2008) ile Rupp ve Templin (2008) tarafından belirtildiği üzere, BTM odaklı bir test oluşturmak için, test maddelerine verilen cevapların nasıl oluştuğunu ve elde edilen verinin yapısını açıklayan kuram veya hipotezleri barındıran bilişsel bir model temel alınmalıdır. Ancak, literatüre bakıldığında, birçok gerçek veri kullanımına bağlı BTM uygulamasının simülasyon çalışmalarına ek olarak ortaya koyulduğu ve halihazırda toplanan verilere sonradan model ekleme (retrofitting) faaliyetlerinin ağırlıkta olduğu görülmektedir (Gierl, Alves ve Majeau, 2010).

Ölçme-değerlendirme süreçlerinde madde/test yanlılığı önemli bir geçerlilik sorunu olarak karşımıza çıkmaktadır (Kane, 2013). Bu sorunla başa çıkmak adına değişen madde fonksiyonu (DMF) tespiti yararlı bir yöntem olarak değerlendirilmektedir. DMF-maddelerini belirlemek için klasik test kuramını (KTK) ve madde tepki kuramını (MTK) temele alan DMF belirleme teknikleri ortaya koyulmuştur. Son zamanlarda, BTM çerçevesinde DMF belirleme teknikleri de literatüre kazandırılmaktadır. Yaygın kullanımı olan BTM’lerden DINA (the deterministic input, noisy "and" gate: Junker & Sijtsma, 2001) modelin veri analizinde kullanıldığı durumlar için Wald testine bağlı olarak DMF belirleme tekniği geliştirilmiştir (Hou, de la Torre ve Nandakumar, 2014). Bu çalışmada, KTK, MTK ve BTM tabanında geliştirilmiş DMF belirleme teknikleri kullanılmış ve sonuçların uyumluluğu değerlendirilmiştir. Özellikle, BTM çerçevesinde geliştirilmemiş olan testlerden elde edilen verilerin sonradan eklenen bir BTM ile analizi sonucunda maddelerin DMF gösterme durumları incelenmiştir. Bu analizlerle testin geliştirilmesinde dikkate alınan ve test sonuçlarının analizinde kullanılan psikometrik modellerin aynı olmadığı durumlarda cinsiyet gibi bağımsız değişkenlerce oluşturulacak alt gruplar için maddelerde DMF görülme durumunun farklılaşıp farklılaşmadığının incelenmesi hedeflenmektedir. Bu çalışmanın ikincil amacı kitapçık türünün psikometrik özellikleri (örneğin madde parametreleri) üzerindeki etkisinin DMF belirleme teknikleri aracılığıyla incelenmesidir. DMF maddelerin psikometrik özelliklerinin alt gruplara göre farklılık göstermesi neticesinde oluştuğundan, test kitapçıklarında (maddelerin sıralaması değiştiğinde) gözlemlenen DMF analiz sonuçlarındaki varyasyon testteki maddelerin sıralamasının kestirilen parametreleri etkilediğine yönelik ampirik kanıt olarak sunulacaktır.

Yöntem

Yukarıda belirtilen hedefler çerçevesinde, bu çalışmada 2013 yılında Ankara ilinde TEOG sınavına girmiş olan 39146 erkek ve 37318 kadın adayın 19 çoktan seçmeli matematik maddesine verdiği cevaplardan seçkisiz örnekleme yöntemi ile oluşturulan örneklemler kullanılmıştır. Verilerin elde edilmesinde kullanılan sınav A ve B kitapçığı olmak üzere sınava giren adaylara sunulmuştur. Bu kitapçıklarda maddelerin sıralaması (konumları) farklılık göstermektedir. Bu testten elde edilen toplam veri setinden, 1000 öğrencinin verisini içeren seçkisiz örnekleme ile 100 tane örneklem oluşturulmuştur. Bu örneklemler, cinsiyete göre yukarıda bahsi geçen üç farklı DMF belirleme tekniği ile analiz edilmiş ve elde edilen istatistikler ilgili istatistiksel dağılımlarla karşılaştırılarak 'kadın ve erkek öğrenciler için maddenin fonksiyonu değişmemektedir' şeklinde ifade edilebilecek yokluk hipotezleri test edilmiştir. Test sonuçları, her bir teknik ve test kitapçığı türü için hipotezin reddedilme oranı olarak rapor edilerek ve ayrıca elde edilen p-değerleri kutu-grafiği olarak karşılaştırılmıştır.

Sonuç ve Tartışma

Yokluk hipotezleri reddedilme oranlarına bakıldığında, farklı DMF tekniklerinde maddelere DMF tanısı konulma oranlarında ciddi farklılıklar gözlemlenmektedir. Öncelikle belirtilmelidir ki Wald teste bağlı olarak DINA model ile veriler analiz edildiğinde ortalama DMF gözlemlenme oranları, sırasıyla α = .05 ve α = .01 anlamlılık düzeylerinde, .47 ve .31 olarak ortaya hesaplanmıştır. Bu haliyle DINA modeli üzerinden Wald test DMF belirleme tekniği en yüksek DMF sonuçlarını doğurmuştur. Mantel-Haenzsel sonuçlarından çok da farklı olmada dahi, Raju'nun alan ölçüleri tekniğiyle DMF analizi yapıldığında elde edilen maddelerde DMF görülme oranlarının ortalaması en düşük seviyede seyretmiştir (α = .05 ve α = .01 olduğunda sırasıyla .28 ve .11). Son olarak, Mantel-Haenzsel DMF belirleme tekniği, α = .05 ve α = .01 altında, sırasıyla, maddelerde .31 ve .16 oranlarında DMF rapor etmiştir. Böylesine bir sonucun olası nedenlerinden biri, orijinal sınavın BTM'ye bağlı olarak geliştirilmemiş olması olarak düşünülebilir. Test tarafından ölçülecek niteliklerin belirlenmesi, nitelik setini ölçen maddelerin geliştirilmesi ve maddeler ile nitelikler arasında doğru bir ilişkinin kurulması için Q-matrisinin oluşturulması, BTM çerçevesinde hazırlanan testten maksimum düzeyde bilgi elde etmek için kilit adımlardır. Bu nedenle, bir testte yer alan maddelerin ve niteliklerin doğru şekilde ilişkilendirilmesi, bilişsel tanıya yönelik değerlendirmenin etkililiğini artırmak için çok önemli bir adım olacaktır. Bu çalışmada kullanılan test ve verilere özgü olmaksızın, genel olarak, bilişsel tanı modellemesi çerçevesinde hazırlanmamış testlerden elde edilen veriler üzerinde sonradan eklenen bir BTM ile analizine yönelik atılacak adımlarda, testin ve test maddelerinin psikometrik özellikleri (örneğin madde parametreleri) hatalı kestirilebilecektir.

DINA model ile yapılan analizler için Wald testine bağlı olarak DMF sonuçları incelendiğinde, kitapçıklar arasında bu tekniğin DMF belirleme oranlarında açık farklılıklar gözlenmiştir. Detaylandırılacak olursa, α = .05 düzeyinde, yedi maddenin DMF gösterme eğilimleri büyük ölçüde farklılaşmıştır. Anlamlılık seviyesi α = .01'e düşürülmüş olsa bile bu yedi maddeden beşi hala belirgin şekilde DMF gösterme eğilimlerinde farklılıklar sergilemişlerdir. Benzer şekilde Raju'nun alan ölçüleri ve Mantel-Haenzsel DMF teknikleri ele alındığında ise dörder maddede DMF gösterme eğiliminde kitapçıklar arasında yüksek farklılıklar ortaya çıkmıştır. Yokluk hipotezlerinin reddedilme oranlarından yola çıkarak yaptığımız değerlendirmede sunulan oranlar analizlerde raporlanan gözlenen p-değerleri sırasıyla .05 ve .01 kritik değerleriyle karşılaştırılarak elde edilmiştir. O halde, yokluk hipotezlerin .051 mi yoksa .999 gibi bir p-değeriyle mi reddedildiği bilinememektedir. Bu nedenle, yokluk hipotezi reddetme oranlarına ek olarak, her bir kitapçık için ele alınan 100 veri setinin analizlerinden elde edilen p-değerleri kutu-grafikleri olarak sunulmuştur ve bu grafikler DMF teknikleri ve kitapçık türleri arasında maddelerde DMF gözlemlenme eğilimlerinin kıyaslanmasında kullanılmıştır.

Kitapçık türlerinden alınan örneklemler üzerinde her üç DMF tekniğiyle cinsiyet grupları açısından maddelerin DMF gösterime eğilimlerinin kutu grafikleriyle incelenmesi sonucunda yukarıda açıklanan bulgularla benzer sonuçlar elde edilmiştir. Dolayısıyla, farklı kitapçıkların maddelerin psikometrik özelliklerinin kestirimi üzerinde bir etkiye sahip olduğu, bir diğer ifadeyle, maddelerin test içerisindeki sıralamalarının maddelerin kestirilen parametrelerine etki ettiğine yönelik ampirik bulgulara ulaşılmıştır. Maddelerin sıralamalarındaki değişikliler farklı alt gruplar için farklı sonuçlar doğurmuş ve dolayısıyla alt gruplar arasında (bu çalışmada cinsiyet grupları arasında) maddenin kestirilen parametrelerinde farklılıklar ortaya çıkmıştır. Alanyazın incelendiğinde, madde konumlarındaki değişikliklerin maddelerin zorluk seviyelerini değiştirdiğini öne süren çalışmalar bulunmaktadır (Kingston ve Dorans, 1984). Bu nedenle, bir testte madde sırası değiştikçe, madde yanıtlama hızında, yanıt oluşturma stratejilerinde, bilişsel çaba harcama oranında ve alt gruplardaki yorgunluk seviyesinde meydana gelebilecek farklılıklar, madde parametrelerinin kestirilen değerlerinde değişikliğe ve dolayısıyla alt gruplar açısından DMF'ye sebebiyet verebilmektedir. Bu bulgular çerçevesinde, maddelerin konumları kitapçıklar arasında değişiklik gösterse dahi, bu konum değişikliklerin DMF'ye ve sonunda test puanları üzerinde ciddi farklılıklara sebep olacak kadar büyük olmaması önem taşımaktadır.