ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 97-111

Retrofitting of Polytomous Cognitive Diagnosis and Multidimensional Item Response Theory Models *

Levent YAKAR** Nuri DOĞAN *** Jimmy DE LA TORRE ****

Abstract
In this study, person parameter recoveries are investigated by retrofitting polytomous attribute cognitive diagnosis and multidimensional item response theory (MIRT) models. The data are generated using two cognitive diagnosis models (i.e., pG-DINA: the polytomous generalized deterministic inputs, noisy "and" gate and fA-M: the fully-additive model) and one MIRT model (i.e., the compensatory two-parameter logistic model). Twenty-five replications are used for each of the 54 conditions resulting from varying the item discrimination index, ratio of simple to complex items, test length, and correlations between skills. The findings are obtained by comparing the person parameter estimates of all three models to the actual parameters used in the data generation. According to the findings, the most accurate estimates are obtained when the fitted models correspond to the generating models. Comparable results are obtained when the fA-M is retrofitted to other data or when the MIRT model is retrofitted to fA-M data. However, the results are poor when the pG-DINA is retrofitted to other data or the MIRT is retrofitted to pG-DINA data. Among the conditions used in the study, test length and item discrimination have the greatest influence on the person parameter estimation accuracy. Variation in the simple to complex item ratio has a notable influence when the MIRT model is used. Although the impact on the person parameter estimation accuracy of the correlation between skills is limited, its effect on MIRT data is more significant.

Key Words: Polytomous attribute cognitive diagnosis models, pG-DINA, fA-M, multidimensional item response theory, retrofitting.

INTRODUCTION

Some of the specific measurement procedures used in education and psychology can be applied to one or more attributes. Scales constructed to measure a single skill may also be applied to another, but high correlations between the skills measured may render the scale insensitive to measuring other skills (Reckase, 2007). Consequently, tests may appear to measure only one main skill. However, if the correlations between measured skills are not too high, the main factor may not suppress other factors, particularly in psychological-based measurements. Thus, multiple skills may be measured intentionally or unintentionally.

Various psychometric approaches can be taken when measuring multiple skills. For example, in item response theory (IRT), unidimensional IRT (UIRT) models can be applied multiple times to measure one skill at a time, whereas multidimensional IRT (MIRT) models can be used to measure more than one skill simultaneously.

* This study is based on the first author's PhD thesis entitled "Retrofitting of cognitive diagnosis and multidimensional item response theory models" [Bilişsel tanı ve çok boyutlu madde tepki kuramı modellerinin karşılıklı uyumlarının incelenmesi] and was written under the supervision of the other authors.

** Assist. Prof. Dr., Kahramanmaraş Sütçü İmam University, Faculty of Education, Kahramanmaraş, Turkey, l_yakar@hotmail.com, ORCID ID: 0000-0001-7856-6926

*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara, Turkey, nuridogan2004@gmail.com, ORCID ID: 0000-0001-6274-2016

**** Professor, The University of Hong Kong, Faculty of Education, Hong Kong, China, j.delatorre@hku.hk, ORCID ID: 0000-0002-0893-3863

To cite this article:
Yakar, L., Doğan, N., & de la Torre, J. (2021). Retrofitting of polytomous cognitive diagnosis and multidimensional item response theory models. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 97-111. doi: 10.21031/epod.778861

Received: 10.08.2020
Accepted: 26.05.2021

Multidimensional Item Response Theory (MIRT)

MIRT models were developed to address the main limitation of UIRT models – they assume a single underlying skill. In contrast, MIRT models can be used when multiple skills interact to determine the probability that an individual will respond correctly to the test items (Ackerman, Geirl & Walker, 2003). These models can produce ability parameter estimates that correspond to the measured skills (Reckase, 2009). MIRT applications have become increasingly common, as test items typically measure more than one skill.

Various MIRT models have been developed and are generally classified as either compensatory or noncompensatory models. In compensatory models, high levels of individual ability in one dimension can make up (i.e., compensate) for lower ability in another dimension. Noncompensatory models are harder to estimate, particularly if exploratory analysis is required (Chalmers & Flora, 2014), and so compensatory models are more commonly used in the field.

MIRT models can also be differentiated based on the number of item parameters involved. If only the item difficulty parameter *d* is involved, the MIRT model will be deemed to belong to the one-parameter model family; for those that belong to two-parameter model family, the item discrimination parameter vector $\boldsymbol{a}$ will be included in addition to *d*; and for those that belong to the three-parameter model family, the pseudo-guessing parameter *c* will be included in addition to $\boldsymbol{a}$ and *d*.

The compensatory two-parameter logistic (2PL) MIRT model introduced by McKinley and Reckase (1982) is widely used. Here, the probability of an individual *i* answering item *j* correctly is given by the formula:

$$p(\boldsymbol{\theta}_i, \boldsymbol{a}_j, d_j) = \frac{1}{1+\exp(-D\sum_{k=1}^{m}(a_{jk}\theta_{ik})+d_j)},$$

where $D = 1.7$ is the measurement constant; θ_{ik} is individual *i*'s *k*th ability parameter; a_{jk} and d_j are the *k*th discrimination parameter and difficulty parameter of item *j*, respectively; and *m* is the number of dimensions.

Cognitive Diagnosis Model (CDM)

Other families of psychometric models called cognitive diagnosis models (CDMs) also available in the pertinent literature. These models were developed to be used in conjunction with cognitively diagnostic assessments (de la Torre & Minchen, 2014). The main purpose of CDM is to determine whether individuals have mastered the attributes or skills measured by the test. As such, CDMs classify individuals based on their mastery profiles, which can be used to identify learning deficiencies. CDM research has recently increased, as CDMs are more effective for measuring finer-grained skills than IRT models (Rupp, Templin & Henson, 2010; von Davier & Lee, 2019).

CDMs classify individuals into latent categories, which are determined by the presence or absence of the measured skills. This classification is based on the individuals' skills, or estimated mastery status, in terms of the measured attributes. The mastery of an attribute is represented by 1, while nonmastery is represented by 0 represents. A correct response to an item signals mastery of the attributes required to correctly respond to the item. A high proportion of correct responses to items requiring a specific attribute may indicate that an individual has already mastered this attribute (Rupp, Templin & Henson, 2010).

The Q-matrix, a common feature of CDMs, is used to define associations between measured attributes and test items. In a Q-matrix, items are placed in rows and attributes in columns. The Q-matrix is essential in a CDM and plays an important role in defining individuals' attribute profiles, as the Q-matrix clarifies the attribute requirements of each item (de la Torre & Minchen, 2014).

CDMs are commonly classified based on whether the measured attributes are dichotomous or polytomous in nature. Dichotomous attributes are those specified as either required (i.e., 1) or not

required (i.e., 0) for correct responses to items in the Q-matrix. Similarly, the attribute profile estimates of individuals are represented by either 0 (i.e., nonmastery) or 1 (i.e., mastery) when the measured attributes are dichotomously scored. If the attributes are polytomously scored, different levels of measured attributes may be required for a successful response to an item, and individuals may have mastered the attributes at different levels. For example, for an attribute with three categories, there may be nonmastery (i.e., 0) along with two mastery levels (i.e., 1, 2). Polytomous attributes may thus reflect different levels of item difficulty associated with the different levels of the measured skills.

Models that consider polytomous attributes can be viewed as extended versions of those that consider dichotomous attributes. These extended models are more flexible and can address problems that generally dichotomous models cannot. Thus, dichotomous models have been generalized to polytomous models. The polytomous G-DINA (pG-DINA: Chen & de la Torre, 2013) model is an example of polytomous CDMs, and is the polytomous version of the *generalized deterministic inputs, noisy "and" gate* model (G-DINA; de la Torre, 2011).

Polytomous Generalized Deterministic-Inputs, Noisy "And" Gate (pG-DINA)

General CDMs can be reduced to specific CDMs by applying restrictions. General, unrestricted models are referred to as saturated, and specific restricted models as reduced (de la Torre & Lee, 2013). For example, the G-DINA is a saturated model, from which several reduced models, such as *deterministic inputs, noisy "and" gate* (DINA; Junker & Sijstma, 2001) and *deterministic inputs, noisy "or" gate* (DINO; Templin & Henson, 2006) can be derived. Similarly, the pG-DINA has been proposed as a saturated model and can be reduced to restricted polytomous models through various constraints.

The pG-DINA first reduces the number of possible attribute vectors into *reduced attribute vectors* by considering only the attributes required by an item. It then further reduces the number of attribute vectors into *collapsed attribute vectors* by only considering the levels of the required attributes. The number of reduced attribute vectors is computed by $M^{K_j^*}$, where M represents the attribute level and K_j^* the number of attributes required by item j. The number of collapsed attribute vectors is equal to the dichotomous G-DINA case and defined by $2^{K_j^*}$. For example, consider an item that measures two of the three $K = 3$ attributes, each with three $M = 3$ levels, as in, 0, 1, and 2. Assume further that this item requires levels 2 and 1 of the first and second attributes, respectively. Thus, the q-vector for this item is (2 1 0). The original, reduced, and collapsed attribute vectors, as defined by Chen and de la Torre (2013), are given in Table 1.

Table 1 shows that $3^{K_j^*} = 3^2 = 9$ reduced attribute vectors are obtained when only considering the attributes that are required by item j. Similarly, the collapsed attribute vectors are obtained by comparing the attribute levels in the reduced attribute vectors to those specified in the q-vector of item j. When an attribute level of the reduced attribute vector is equal to or higher than the level specified in the q-vector, it is represented by 1 in the collapsed attribute vector. Otherwise, it is represented by 0. The number of collapsed attribute vectors in this example then reduces to $2^{K_j^*} = 2^2 = 4$.

Table 1. Reduced and Collapsed Attribute Vectors for Original Attribute Vectors

Original $\boldsymbol{\alpha}_{lj}$	Reduced Attribute Vector ($\boldsymbol{\alpha}_{lj}^{*}$)	Collapsed Attribute Vector ($\boldsymbol{\alpha}_{lj}^{**}$)
(0,0,0), (0,0,1), (0,0,2)	(0,0)	(0,0)
(1,0,0), (1,0,1), (1,0,2)	(1,0)	
(0,1,0), (0,1,1), (0,1,2)	(0,1)	(0,1)
(1,1,0), (1,1,1), (1,1,2)	(1,1)	
(0,2,0), (0,2,1), (0,2,2)	(0,2)	
(1,2,0), (1,2,1), (1,2,2)	(1,2)	
(2,0,0), (2,0,1), (2,0,2)	(2,0)	(1,0)
(2,1,0), (2,1,1), (2,1,2)	(2,1)	(1,1)
(2,2,0), (2,2,1), (2,2,2)	(2,2)	

The probability of success associated with the collapsed attribute vector or latent group $\boldsymbol{\alpha}_{lj}^{**}$ computed using the pG-DINA function is as follows:

$$P(\boldsymbol{\alpha}_{lj}^{**}) = \delta_{jo} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk}^{**} + \sum_{k'>k}^{K_j^*}\sum_{k=1}^{K_j^*} \delta_{jkk'}\alpha_{lk}^{**}\alpha_{lk'}^{**} + \cdots + \delta_{j1,\ldots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lK_j^*}^{**}.$$

The interpretations of the model parameters are the same as those for the dichotomous attribute cases in the G-DINA model. Whereas the pG-DINA model uses the collapsed attribute vectors given in Table 1, the fully additive model (fA-M; Yakar, de la Torre, & Ma, 2017), another polytomous CDM, considers the reduced attribute vectors.

Fully Additive Model (fA-M)

If restrictions are applied to the saturated pG-DINA model, it can be reduced to a polytomous additive CDM (pA-CDM; Chen & de la Torre, 2013). This pA-CDM is derived from the pG-DINA by setting all interaction effects to zero. The intercept and the main effects of the mastered attributes required by the item are summed in the pA-CDM to obtain the probability of a correct response to the item. The fA-M can also be considered as an additive and restricted model. The main difference between these two models is the latent classes for which they compute the item response functions. The pA-CDM only considers the collapsed attribute vectors like the pG-DINA model, whereas the fA-M considers reduced attribute vectors.

Although fA-M is a restricted model, incorporating the reduced rather than the collapsed latent class in the item response function distinguishes it from many other CDMs. Rather than all-or-none, the fA-M considers the contributions of all levels (i.e., 0, 1, 2,…), as in, it considers the levels of the polytomous attribute in computing the probability of a correct response. This characteristic indicates that the model mimics the compensatory MIRT model, as the higher level of skills (i.e., attributes) leads to a higher probability of a correct response. The item response function of fA-M is given as

$$P(\boldsymbol{\alpha}_{lj}^{*}) = \delta_{j0} + \sum_{k=1}^{K_j^*}\sum_{m=1}^{M_k} \delta_{jkm}\alpha_{lk}^{*},$$

where δ_{j0} is the intercept, δ_{jkm} is the (main) effect of the m^{th} level of attribute k, K_j^* is the number of required attributes, and M_k is the highest level of attribute k.

A characteristic common to CDMs and MIRT models is that both can be used with multidimensional scales. In addition, both theories contain compensatory and noncompensatory models (Reckase, 2009; Rupp, Templin, & Henson, 2010). These similarities indicate that these model families can be used to

estimate multiple attributes or abilities. The type of the item structure in these models are also common, as they can be simple or complex in both CDMs and MIRT models, which is of particular importance in the analyses. However, these families of models differ in terms of other features, such as item parameters, the nature of the person parameters, which can be continuous or discrete, and the measurement units used.

The similarities between these psychometric models imply that deciding which model to use can be an issue of high consideration. Under some analysis conditions, fitting various models to the data may provide different points of view and lead to a deeper understanding – comparing the results obtained from different models that have similar infrastructures can extend our understanding of the focal phenomenon. Thus, evaluating the outputs of CDMs and MIRT models together can be of value.

To obtain additional information, a model that does not share psychometric properties with the tests used to gather the data may be fitted to the data. This process, referred to as retrofitting, and can be used to obtain potentially different information that supports or refutes existing knowledge about the data. The results of a retrofitting analysis may be much more valuable when the true and retrofitted models have similar structures, as with the CDM and MIRT models.

A literature review reveals that many studies have focused on retrofitting CDM to IRT data, and vice versa. Various CDMs are retrofitted to data obtained via tests that have been developed for IRT purposes (Ardıç, 2020; Chen & Chen, 2016; Chen & de la Torre, 2014; Lee, Park & Taylan 2011; Liu, Huggins-Manley, & Bulut, 2018; Şen & Arıcan, 2015). Other studies (de la Torre & Karelitz, 2009; Wang, 2009) involve reciprocal retrofitting CDMs and MIRT models. However, no retrofitting study that focuses on polytomous CDMs has been identified. Therefore, a significant contribution of the current study is the reciprocal retrofitting of three models: two CDMs and one MIRT model.

Purpose of the Study

The aim of this research was to examine the level of information obtained through retrofitting two specific CDMs and a MIRT model. We addressed this through the following sub-problems:

1- What levels of accuracy can be obtained for the person parameter classification and ability level estimation from the two CDMs and one MIRT model when they are fitted to the MIRT data generated under various item discrimination, item structure, correlation between skills, and test length conditions? Is there a difference between the person parameter estimation accuracy levels of the models?
2- What levels of accuracy can be obtained for the person parameter classification and ability level estimation from the two CDMs and one MIRT model when they are fitted to the CDMs data generated under various item discrimination, item structure, correlation between skills, and test length conditions? Is there a difference between the person parameter estimation accuracy levels of the models?

METHOD

Research Type

Experimental or theoretical studies that do not have any apparent specific application or use, and are primarily carried out to obtain novel information on the basis of phenomena and observable facts are defined as basic research (OECD, 2002). This study can be considered basic research as the aim is to assess the comparability of the results from fitting a MIRT model and two CDMs to various data. The data were generated using the models considered, and the analytic performance of the retrofitted models and the generating models were then examined.

Data Generation

Item discrimination, item structure, test length, and correlation between skills were manipulated to obtain various conditions simulation conditions. Three levels of item discrimination were specified and the generated discrimination parameters were drawn from uniform distributions, as in, $a \sim U(0.6, 0.8)$, $U(0.9,1.1)$, and $U(1.5,1.7)$ for the low, moderate, and high item discrimination conditions, respectively. The item structure was defined in terms of item complexity (i.e., whether the item measures one or more dimensions/attributes). In this research, an item is said to have a simple structure if it measures only one dimension/attribute, and a complex structure, otherwise. Tests with Q-matrices consisted of 20%, 50%, and 80% simple structure items were considered to have mostly complex, equal, and mostly simple item structures, respectively. In terms of the test length condition, the three levels of test length (i.e., short, medium, and long) consisted of 15, 30, and 60 items, respectively. The two levels of correlation (i.e., no relationship and moderate relationship) were created by setting the correlation between skills to .00 and .60. Although the correlation cannot be zero in real data cases and under compensatory models, this value was nonetheless considered because it reflects a situation in which a relationship is not present, which may provide a better understanding of the parameters in its related state. In terms of factor selection and their levels, the conditions used in other similar studies (Chen & de la Torre, 2013; Wang, 2009) and factors affecting model performance were considered. The study was conducted with 25 replications, as analyzing polytomous attribute data takes longer than when using dichotomous attribute data (de la Torre & Douglas, 2004; de la Torre & Douglas, 2008; Huebner & Wang, 2011). Thus, in the three models, three-item discrimination levels, three-item structure levels, three test length levels, and two correlation levels were crossed to yield $3 \times 3 \times 3 \times 3 \times 2 = 162$ conditions. With 25 replications for each condition, a total of $162 \times 25 = 4050$ data were generated and analyzed using the two CDMs and the MIRT model.

Generation of MIRT data

For each level of crossed factors, two-dimensional 2PL MIRT data were generated using the R program. For this data generation, the ability parameters followed a multivariate normal distribution, and the attribute levels in the polytomous CDMs indicated the item difficulty levels. Specifically, the Q-matrix entries of the polytomous CDMs were transformed into the item difficulty parameters. When generating the data, the item difficulty parameters were obtained by multiplying each element of the Q-matrices by 0.67 and subtracting 1.34 from each. Accordingly, the levels of 0, 1, 2, and 3 in the Q-matrix correspond to the difficulty levels of -1.34, -0.67, 0, and 0.67, respectively. As 0 in a Q matrix stands for an unmeasured attribute, the discrimination parameters of the items with the difficulty parameter of -1.34 were set to zero to ensure that the item parameters of the CDMs and the MIRT model were as matched as possible.

The continuous person parameters in MIRT were converted to discrete attribute levels in order to obtain classification accuracy rates that can be compared. By applying the cut-off points (i.e., -0.67, 0, and 0.67) to the MIRT person parameters, discrete values of 0, 1, 2, and 3 were obtained for each dimension resulting in individuals being classified into approximately four equal groups for each dimension. The sample size was set to 5000 to obtain more stable item and person parameter estimates.

Generation of CDM data

Two CDMs were used in this study: fA-M and pG-DINA models. As the item parameters of these CDMs differed in terms of number and structure, a two-dimensional 2PL MIRT data of 100000 examinees was initially generated to obtain related conditions for the CDMs. Item parameters compatible with the fA-M and pG-DINA model were then obtained in the R environment for two attributes. A self-written R code and the GDINA package (Ma & de la Torre, 2016) were used to generate the data for the fA-M and pG-DINA model, respectively.

Data Analysis

The generated data in MIRT were analyzed using the MIRT package (Chalmers, 2012). Person parameter estimates were obtained based on the expected a posteriori (EAP) method. The estimated person parameters were converted into discrete variables, similar to the generated person parameters in order to obtain the classification accuracy rates of the person parameter estimates under MIRT conditions. Analyses of MIRT data in CDM were performed using the GDINA package (Ma & de la Torre, 2016) for pG-DINA cases and through a self-written R code for fA-M cases.

Although the data in the MIRT estimation of the CDM data were originally based on 2PL, the relative fit of 2PL and 3PL MIRT models were both checked. After the data were analyzed in both models, ANOVA tests on deviance indices were conducted through R. If the difference was statistically significant ($p < .05$), the parameters of the 3PL model were considered. In general, the 3PL model was observed to fit better with the data.

After the analyses, the correct vector classification rates (CVCR) of the person parameters were obtained. If the estimated and generating ability/attribute vectors of an examinee matched, the examinee was considered to be accurately classified by the estimating model. The ratio of the number of accurately classified examinees to the total number in a dataset (i.e., 5000) provides the CVCR of the model. The average CVCR of the study was obtained across 25 replications. The significance of the differences between the CVCRs across models was tested through ANOVA. Since violation of the equality of variance assumptions, pairwise comparisons of groups were performed using the Tamhane procedure.

The ability/attribute-level accurate classification rates reflect the degree to which each dimension/attribute level is accurately estimated by the model. It, therefore, reflects the performance of the model at the individual ability/attribute levels. Accordingly, the averages of the correct attribute level classification rates (CALCRs) of the ability/attribute levels of all examinees on two abilities/attributes were obtained.

Data were generated for each model under 54 different conditions by crossing the main factors. As these factors are independent of each other, no interaction between different conditions was identified; thus, the CVCR averages at different levels of the basic conditions were reported rather than at the level of the crossed conditions. The findings can thus be effectively presented and interpreted. The CVCR averages across the conditions and repetitions are presented in Appendix.

RESULTS

This section presents the results of retrofitting the CDMs to the MIRT data, as stated in the first sub-problem, and of retrofitting the MIRT model to the CDM data, as stated in the second sub-problem.

Results of the MIRT Data Analysis

The CVCRs obtained by analyzing the MIRT data are presented in Table 2. The table shows that the highest CVCRs are obtained for the MIRT data when the fitted model was the MIRT model, followed by the fA-M. The CVCRs ranged from .41 to .60 when the MIRT model (i.e., the generating model) was fitted, and from .36 to .52 when the fA-M model was retrofitted to the data. The lowest levels of correct classification rates were observed when the pG-DINA model was retrofitted to the data, where the CVCRs ranged between .26 and .33. These findings suggest that the CVCRs of the MIRT analyses and the fA-M retrofitting results were comparable, which were different from the CVCRs of the pG-DINA analyses.

Table 2. CVCRs Obtained from the MIRT Data

Condition	Level	MIRT	pG-DINA	fA-M
Item discrimination	Low	0.43	0.27	0.40
	Moderate	0.50	0.29	0.45
	High	0.58	0.33	0.48
Item structure	Mostly complex	0.45	0.27	0.38
	Equal	0.52	0.29	0.43
	Mostly simple	0.54	0.33	0.51
Test length	15	0.41	0.28	0.36
	30	0.50	0.30	0.44
	60	0.60	0.32	0.52
Correlation between abilities	0.00	0.48	0.26	0.38
	0.60	0.53	0.33	0.50

As the item discrimination increased, the correct classification rates of all three models also increased. Moving from lower to higher discrimination levels, the increment for the correct classification performance of the MIRT analyses (.15) was larger than those of the CDM cases (.06 for pG-DINA and .08 for fA-M model). Similarly, regardless of the models, higher CVCRs were observed under mostly simple item conditions. The average increments in the CVCRs of the MIRT, pG-DINA, and fA-M model conditions were .09, .06, and .13, respectively.

In terms of the test length condition, an increase in the test length improved the CVCRs – the mean CVCRs increased from .41 to .60 and from .36 to .52, respectively, when the MIRT model and fA-M were fitted to the data. A relatively smaller increase (i.e., from .28 to .32) was observed when the fitted model was the pG-DINA.

The CVCRs also tended to increase when the abilities were correlated. This increase was larger for the retrofitted CDMs, particularly for the fA-M.

The ANOVA test results presented in Table 3 demonstrate the differences among the CVCRs of all three models when they were fitted to the MIRT data. The results indicate a significant difference between the CVCRs obtained through the analysis of the MIRT data [$F(2,4047) = 1592.984$, $p < .001$]. A pairwise comparison of the CVCRs obtained from the models using the Tamhane method reveals that the CVCR of the MIRT is significantly higher than those of the CDMs ($p < .001$). Similarly, the CVCR of the fA-M is significantly larger than that of the pG-DINA model ($p < .001$).

Table 3. Test of the Difference Between CVCRs of the MIRT Data Analyses

Variance Source	Sum of Squares	df	*F*	Difference
Between groups	30.807	2	1592.984*	MIRT>fA-M
Within group	39.132	4047		MIRT>pG-DINA
Total	69.939	4049		fA-M>pG-DINA

*$p<.001$

The attribute-level correct classification rates of all three models fitted to the MIRT data are presented in Table 4. The most significant results observed for the pG-DINA models were that the attribute-level CALCRs for levels 0 and 3 were very high (i.e., .96), but the CALCRs of levels 1 and 2 were very low (i.e., .10). Although the CALCRs in attribute levels 0 and 3 were also higher than those in attribute levels 1 and 2 when the MIRT model and fA-M were fitted to the data, the difference was not as dramatic. For attribute levels 0 and 3, the MIRT and fA-M CALCRs are .79 and .74, respectively; the corresponding CALCRs for attribute levels 1 and 2 are .61 and .56, respectively.

Table 4. CALCRs in the Analyses of MIRT Data

Model	Level 0	Level 1	Level 2	Level 3
MIRT	0.79	0.61	0.61	0.79
pG-DINA	0.96	0.10	0.10	0.96
fA-M	0.74	0.56	0.56	0.74

Results of the pG-DINA Data Analysis

The CVCRs obtained from the analysis of the pG-DINA data are presented in Table 5. The table shows that the largest CVCRs are obtained for the pG-DINA data when the fitted model was the generating model (i.e., pG-DINA model), followed by the fA-M. The CVCRs vary from .53 to .90 when the pG-DINA model (i.e., the generating model) was fitted, and from .51 to .89 when the fA-M model was fitted to the data. The lowest correct classification rates are observed when the MIRT model was retrofitted to the data – the CVCRs vary between .31 and .56. These findings suggest that the CVCRs of the pG-DINA model and the fA-M are comparable (i.e., the maximum difference is .02), whereas those of MIRT were quite different.

The correct classification rates for all three models increased with the item discrimination. Moving from lower to higher discrimination levels, the increment for the correct classification performance of the MIRT analyses (.19) was slightly lower than that of the pG-DINA model and fA-M (i.e., .24 and .23, respectively). When the items became simpler, an apparent increase in the CVCRs of MIRT was observed (i.e., .21), whereas at most, a slight increase (i.e., .02 in pG-DINA cases) was observed when CDMs were used in the data analysis. In terms of the test length, the CVCRs increased with the test length – the mean CVCRs improved from .53 to .90 and from .51 to .89 when the pG-DINA model and fA-M were fitted to the data, respectively. A relatively smaller increase (i.e., from .32 to .56) was observed when the fitted model was the MIRT. In addition, the CVCRs also increased when the abilities were correlated, although only to a very limited extent.

Table 5. CVCRs Obtained in the Analyses of pG-DINA Data

Condition	Level	pG-DINA	MIRT	fA-M
Item Discrimination	Low	0.60	0.35	0.59
	Moderate	0.72	0.44	0.72
	High	0.84	0.54	0.82
Item Structure	Mostly Complex	0.73	0.31	0.70
	Equal	0.72	0.50	0.71
	Mostly Simple	0.71	0.52	0.70
Test Length	15	0.53	0.32	0.51
	30	0.73	0.45	0.72
	60	0.90	0.56	0.89
Correlation between Abilities	0	0.71	0.44	0.70
	0.6	0.73	0.45	0.72

Table 6 displays the ANOVA test results of the observed differences between the CVCRs of all three models when they were fitted to the pG-DINA data. The results indicate a significant difference between the CVCRs obtained from the pG-DINA data analysis [$F(2,4009) = 1010.622$, $p < .001$]. A pairwise comparison of the CVCRs obtained from the models using the Tamhane method revealed that the CVCR of the MIRT was significantly lower than those of the CDMs ($p < .001$). In addition, the CVCRs of the fA-M were not significantly different from those of the pG-DINA model ($p > .05$).

Table 6. Test of the Difference between CVCRs of the pG-DINA Data Analyses

Variance Source	Sum of Squares	df	*F*	Difference
Between groups	65.273	2	1010.622*	pG-DINA> MIRT
Within group	129.464	4009		fA-M > MIRT
Total	194.737	4011		

*p<.001

Table 7 presents the attribute-level correct classification rates of all three models when they are fitted to the pG-DINA data. CALCRs of pG-DINA model and fA-M were significantly larger than the CALCRs of the MIRT model. Although the CALCRs in attribute levels 1 and 2 were lower than those in attribute levels 0 and 3 (i.e., the smallest is .81 and the largest .87), the largest CALCRs were obtained when the fitted model was the generating model (i.e., pG-DINA). These were followed by the CALCRs obtained when the fA-M was fitted, which is more uniform across attribute levels (i.e., the smallest is .80 and the largest .81). The lowest CALCRs were observed when the fitting model was MIRT, and the lowest and highest are .56 and .71, respectively.

Table 7. CALCRs in the Analyses of pG-DINA Data

Model	Level 0	Level 1	Level 2	Level 3
pG-DINA	0.87	0.81	0.80	0.87
MIRT	0.65	0.56	0.60	0.71
fA-M	0.81	0.80	0.80	0.81

Results of the fA-M Data Analysis

The CVCRs obtained from the analysis of the fA-M data are presented in Table 8. The table shows that the largest CVCRs were obtained for the fA-M data when the fA-M was fitted, and these varied from .44 to .80. The minimum and maximum CVCRs of the MIRT model, when it was retrofitted to the fA-M data under various conditions, were .39 and .71, respectively. The lowest correct classification rates were observed when the pG-DINA model was fitted to the data – the CVCRs varied between .24 and .34. The CVCRs of the MIRT and fA-M models were comparable; however, the performance of pG-DINA model was relatively poor.

Table 8. CVCRs Obtained in the Analysis of fA-M Data

Condition	Level	fA-M	MIRT	pG-DINA
Item Discrimination	Low	0.50	0.46	0.26
	Moderate	0.61	0.55	0.29
	High	0.74	0.65	0.34
Item Structure	Mostly Complex	0.59	0.47	0.28
	Equal	0.63	0.57	0.29
	Mostly Simple	0.63	0.61	0.31
Test Length	15	0.44	0.39	0.27
	30	0.61	0.55	0.29
	60	0.80	0.71	0.32
Correlation between Abilities	0	0.61	0.55	0.30
	0.6	0.62	0.55	0.29

In terms of the effects of the examined factors on CVCRs, the correct classification rates of all three models increased with the item discrimination. This increment was largest for the fA-M (i.e., .24), followed by the MIRT model (i.e., .19), and the smallest increment was for the pG-DINA model (i.e., .08). Similarly, regardless of the models, higher CVCRs were observed as the number of simple items in a test increased. The average increment for the CVCRs of the MIRT model (i.e., .14) model was relatively higher than the increments observed under the pG-DINA model (i.e., .03) and fA-M (i.e., .04).

In terms of the test length, an increase in the test length resulted in a rise in the CVCRs. The observed increments in CVCRs were very large in the fA-M and MIRT cases, which increased from .44 to .80 and from .39 to .71 when the fA-M and MIRT models were fitted, respectively. However, the increase for the pG-DINA model was limited (i.e., from .27 to .32). In addition, no remarkable changes in CVCRs were observed when the skills/attributes were correlated.

Table 9 presents the ANOVA test results of the observed differences between the CVCRs of all three models when they were fitted to the fA-M data. The table shows a significant difference between the CVCRs obtained from the analysis of the fA-M data [$F(2.4047) = 1934.53$, $p < .001$]. A pairwise comparison of the CVCRs obtained from the models using the Tamhane method revealed that the CVCR of the fA-M was significantly higher than those of the pG-DINA and MIRT models ($p < .001$). Similarly, the CVCR of the MIRT was significantly higher than that of the pG-DINA model ($p < .001$).

Table 9. Test of the Difference between CVCRs of the fA-M Data Analyses

Variance Source	Sum of Squares	df	*F*	Difference
Between groups	77.937	2	1934.53*	fA-M>pG-DINA
Within groups	81.521	4047		fA-M> MIRT
Total	159.458	4049		MIRT>pG-DINA

*p<.001

Table 10 presents the attribute-level correct classification rates of all three models when they were fitted to the fA-M data. The results in this table are comparable to those for the MIRT data given in Table 4. In the pG-DINA cases, CALCRs for levels 0 and 3 were very high (i.e., .98), whereas the CALCRs for levels 1 and 2 were very low (i.e., .11). Although the CALCRs for attribute levels 0 and 3 were also higher than those for levels 1 and 2 when the fA-M and MIRT model were fitted to the data, the differences were not as dramatic for attribute levels 0 and 3, the MIRT and fA-M CALCRs were .82 and .85, respectively, whereas the corresponding CALCRs for levels 1 and 2 were .63 and .68, respectively.

Table 10. CALCRs in the Analyses of fA-M Data

Model	Level 0	Level 1	Level 2	Level 3
fA-M	0.85	0.67	0.68	0.84
MIRT	0.82	0.62	0.63	0.81
pG-DINA	0.98	0.11	0.11	0.98

DISCUSSON and CONCLUSION

This study aimed to examine how the MIRT model and polytomous CDMs, the pG-DINA model (Chen & de la Torre, 2013) and the fA-M (Yakar, de la Torre, & Ma, 2017) when retrofitted to data generated with different underlying processes. Data for each model were generated with varying item discrimination, item structure, test length, and correlation between ability conditions. The data were then fitted with all three models, and the CVCRs of the generated person parameters were examined.

For the first sub-problem, the data generated using the MIRT model were, as expected, most accurately estimated by the MIRT; the fA-M estimation was the next best, and the lowest performance was observed for the pG-DINA model. The results from MIRT and fA-M can be explained due to the use of reduced latent groups in fA-M, which follows a similar logic as the MIRT – a higher level of proficiency corresponds to a higher probability of answering the item correctly. The pG-DINA model is processed through the collapsed latent groups, and an increase in attribute level does not always produce an increase in the success probability, unlike in the fA-M, where every increase in attribute level results in an increase in the success probability.

The highest level of efficiency was obtained from the test length condition in the MIRT data analysis, followed by item discrimination and item structure; the effect of the correlation between abilities was limited compared to other factors. However, the item structure was only effective in the MIRT data or estimation. In addition, the findings revealed that the effect sizes of the conditions may differ in the CDMs. The pG-DINA was less sensitive to changes in the conditions, and the fA-M was more affected by item structure and correlations between skills than the MIRT model. Wang (2009) conducted a reciprocal retrofitting of the reduced reparametrized unified model (R-RUM; DiBello, Roussos, & Stout, 2007; Hartz, 2002) and the MIRT model, and found the estimation accuracy varied according to the item structure and item discrimination. This is consistent with the fA-M results found in the present study. In a different study, where reciprocal retrofitting of one-dimensional IRT and DINA models were examined, de la Torre and Karelitz (2009) found that item discrimination greatly affected the estimation accuracy. Again this is similar to the fA-M results. These results suggest that common factors may affect the performance of different but compatible models in situations involving reciprocal retrofitting.

For the second sub-problem, the analysis of the pG-DINA data indicated that the pG-DINA accurately estimated its own data. The accuracy rates were the highest obtained in the study. The rates obtained from the fA-M were very close to the pG-DINA, and no statistical difference between the results was found. The similar CVCRs of the fA-M and the pG-DINA model when fitted to pG-DINA data is remarkable and provided the best retrofitting results; however, the CVCRs for the MIRT were substantially lower than those of the two CDMs. This finding is consistent with outcomes for the first research problem and suggests that the MIRT model cannot be retrofitted to pG-DINA data, and vice versa.

Although the outcomes were not identical, the successful estimation of the pG-DINA data when fitted with the fA-M may be due to the interaction effects in pG-DINA being substituted with the main effects for each level. The models do not need to have exactly the same item parametrization to produce similar results as different parameters can adjust and fill the gaps when changes in model parametrization occur. To this end, models that contain more item parameters can be more flexible and advantageous. Although the CVCRs of the MIRT were relatively poor for the pG-DINA data, these results were close to the values obtained when fitted to its own data. Thus, at least for the current setup, the MIRT model may not be expected to provide good classification results.

The results of the pG-DINA data analysis revealed that longer test length and higher item discrimination improved the CVCRs of all of the models. In addition, simplifying the item structure resulted in an increase in CVCR of the MIRT model only. Another remarkable result is that the item structure may have limited impact when CDMs are fitted to pG-DINA data.

In the analysis of the fA-M data, it was found that, as in other models, the classification rate was best in the correct model fitted to the data. However, the results of the MIRT model were almost at the same level as those of the fA-M. Similar results were found in the MIRT data analysis. This suggests that the MIRT model and fA-M may be used interchangeably in situations similar to those examined in the current study. As in the MIRT data analysis, the pG-DINA results were again found to have the lowest rates. In terms of the factors considered, all except the correlation between the dimensions had a substantial impact on the CVCRs.

The relatively low CVCRs of the pG-DINA when retrofitted to MIRT and fA-M data were due mainly to the very low CALCRs observed in the two middle attribute levels. A closer inspection (not

presented) showed the pG-DINA model had a tendency to misclassify middle attribute levels as extreme attribute levels. In the study, we discretized the continuous abilities to create a uniform distribution. Poor retrofit performance may be worsened if the abilities have a normal distribution. The poor performance of the retrofitted pG-DINA model may be due to the assumption invoked by the model to create the collapsed latent classes.

When fitting the correct model to the data, the MIRT was found to have lower CVCRs than the pG-DINA model or the fA-M. The poor results suggest that estimating MIRT data may be more challenging. Moreover, the original person parameters of the MIRT are continuous but were made discretized for comparison purposes. The loss of information due to this transformation may have negatively affected the results.

It is worth noting that the CVCRs obtained in retrofitting the fA-M to the pG-DINA data were unexpectedly higher than those obtained in fitting the model to its own data. A similar situation was found for MIRT model – the MIRT estimations of the MIRT data were less accurate than those of the fA-M data. In their retrofitting studies, de la Torre and Karelitz (2009) and Wang (2009) found similar results. These results suggest that the underlying processes in generating the different data vary in complexity. To the extent that the findings can be generalized, the underlying process of the pG-DINA model is the simplest, followed by the fA-M, and the MIRT model has the most complex underlying process.

Overall, fitting a model that corresponds to the true underlying process produced the best results, whereas fitting a wrong model can lead to slightly or substantially poorer results depending on the extent of the mismatch. Of the three models, the fA-M was relatively robust to the possible mismatch between the true and fitted models; the same cannot be said of the pG-DINA and MIRT models. Although model-data fit still needs to be evaluated, fitting the fA-M to real data appears to be a safer option.

A limitation of the current study pertains to how the abilities were converted to attributes. Although the fA-M can be used to extract diagnostic information for polytomous attributes from MIRT data, and vice versa, these results may only be true when abilities are discretized in a particular manner. Future studies should consider other ways of establishing the comparability of the MIRT model and fA-M in order to arrive at more general conclusions. It should be noted that this study does not suggest that the MIRT model and fA-M can be used interchangeably – as pointed out repeatedly, fitting the true model will always produce the best results. To this end, further studies are needed to establish which procedures can be used to identify the best model when inferences about polytomous attributes are of interest.

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-51.

Ardıç, E. Ö. (2020). *Bilişsel tanı ve çok boyutlu madde tepki modellerinin sınıflama doğruluğu ve parametrelerinin karşılaştırılması [Comparison of classification accuracy and parameters of cognitive diagnostic and multidimensional item response models]*. Unpublished PhD Dissertation, Hacettepe University, Ankara.

Chalmers, R. P., (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chalmers, R. P., & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement, 38*(5), 339-358.

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly, 13*(3), 218-230.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419-437.

Chen, J., & de la Torre, J. (2014). A procedure for diagnostically modeling extant large-scale assessment data: The case of the programme for international student assessment in reading. *Psychology, 5*(18), 1967-1978.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595-624.

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46*(4), 450-469.

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*(4), 355-373.

de la Torre, J., & Minchen N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*(2), 89-97.

DiBello, L.V. Roussos L. A., & Stout, W. (2007). *Review of cognitively diagnostic assessment and a summary of psychometric models*. Rao, C. Sinharay, S. (Eds.) Handbook of Statistics, Psychometrics. Vol. 26. North-Holland: Amsterdam.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*, Unpublished PhD dissertation, University of Illinois at Urbana-Champaign.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407-419.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*(2), 144-177.

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*(3), 357-383.

Ma, W. & de la Torre, J. (2016). GDINA: The generalized DINA model framework. R package version 0.9.2.

McKinley R. L. & Reckase M. D. (1982) *The use of the general Rasch model with multidimensional item response data* (Research Report: ONR 82-1). American College Testing, Iowa City, IA.

Organisation for Economic Co-operation and Development. (2002). Frascati Kılavuzu. Paris: OECD.

Reckase, M. D. (2007). Multidimensional item response theory. Rao, C. Sinharay, S. (Ed.) *Handbook of Statistics*, Psychometrics. Vol. 26. North-Holland: Amsterdam.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

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

Şen, S., & Arıcan, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology, 6*(2), 238-253.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

von Davier, M., & Lee, Y. S. (2019). Introduction: From latent classes to cognitive diagnostic models. In *Handbook of Diagnostic Classification Models* (pp. 1-17). Springer, Cham.

Wang, Y. C. (2009). *Factor analytic models and cognitive diagnostic models: How comparable are they? – A Comparison of R-RUM and compensatory MIRT model with respect to cognitive feedback.* Unpublished PhD dissertation, The Faculty of The Graduate School at The University of North Carolina at Greensboro).

Yakar, L., de la Torre, J., & Ma, W. (2017). *An empirical comparison of two cognitive diagnosis models for polytomous attributes*. In the Annual Meeting of National Council on Measurement in Education. National Council on Measurement in Education (NCME), San Antonio, TX.

Appendix. CVCR Averages for Crossed Conditions

Conditions				True Model	Retrofitted Model		True Model	Retrofitted Models		True Model	Retrofitted Models	
Correlation	Test Length	Item Structure	Item Disc.	MIRT	pG-DINA	fA-M	pG-DINA	MIRT	fA-M	fA-M	MIRT	pG-DINA
0	15	M. Complex	Low	.26	.19	.24	.39	.21	.37	.32	.28	.23
0	15	M. Complex	Moderate	.31	.20	.26	.51	.26	.48	.40	.33	.26
0	15	M. Complex	High	.39	.21	.26	.63	.29	.60	.49	.39	.30
0	15	Equal	Low	.31	.21	.27	.38	.24	.37	.34	.31	.23
0	15	Equal	Moderate	.38	.23	.29	.51	.31	.49	.43	.39	.26
0	15	Equal	High	.47	.25	.27	.66	.43	.63	.55	.49	.30
0	15	M. Basic	Low	.32	.23	.31	.39	.26	.38	.35	.33	.23
0	15	M. Basic	Moderate	.41	.27	.37	.51	.37	.50	.44	.43	.28
0	15	M. Basic	High	.52	.33	.42	.66	.53	.65	.57	.55	.35
0	30	M. Complex	Low	.34	.22	.29	.58	.27	.57	.45	.40	.25
0	30	M. Complex	Moderate	.40	.22	.31	.72	.31	.71	.57	.46	.28
0	30	M. Complex	High	.47	.23	.28	.85	.36	.83	.69	.54	.32
0	30	Equal	Low	.40	.24	.35	.58	.36	.56	.50	.47	.26
0	30	Equal	Moderate	.49	.25	.38	.72	.47	.71	.62	.56	.28
0	30	Equal	High	.59	.28	.33	.86	.63	.85	.76	.68	.33
0	30	M. Basic	Low	.43	.25	.40	.56	.38	.54	.49	.49	.26
0	30	M. Basic	Moderate	.52	.28	.47	.71	.51	.71	.62	.60	.28
0	30	M. Basic	High	.63	.33	.59	.86	.73	.86	.77	.75	.37
0	60	M. Complex	Low	.43	.23	.36	.81	.32	.80	.63	.53	.27
0	60	M. Complex	Moderate	.50	.23	.37	.92	.35	.91	.76	.61	.31
0	60	M. Complex	High	.59	.24	.35	.98	.45	.97	.88	.71	.34
0	60	Equal	Low	.52	.25	.43	.80	.52	.80	.68	.64	.28
0	60	Equal	Moderate	.60	.27	.48	.91	.67	.91	.81	.74	.31
0	60	Equal	High	.69	.30	.40	.97	.82	.97	.92	.84	.37
0	60	M. Basic	Low	.55	.28	.48	.77	.56	.76	.68	.66	.29
0	60	M. Basic	Moderate	.64	.31	.58	.90	.73	.89	.82	.79	.33
0	60	M. Basic	High	.72	.43	.69	.97	.60	.97	.93	.89	.42
0.6	15	M. Complex	Low	.39	.29	.38	.43	.23	.42	.33	.29	.23
0.6	15	M. Complex	Moderate	.43	.31	.41	.55	.25	.54	.41	.33	.26
0.6	15	M. Complex	High	.48	.33	.44	.66	.29	.66	.51	.39	.29
0.6	15	Equal	Low	.39	.30	.38	.41	.25	.41	.35	.32	.24
0.6	15	Equal	Moderate	.45	.31	.43	.53	.31	.52	.44	.40	.26
0.6	15	Equal	High	.52	.34	.48	.68	.43	.66	.57	.49	.30
0.6	15	M. Basic	Low	.39	.30	.38	.40	.26	.39	.35	.33	.24
0.6	15	M. Basic	Moderate	.45	.33	.44	.52	.37	.52	.45	.43	.28
0.6	15	M. Basic	High	.55	.37	.53	.68	.54	.67	.58	.56	.34
0.6	30	M. Complex	Low	.45	.30	.44	.65	.27	.64	.47	.40	.26
0.6	30	M. Complex	Moderate	.49	.31	.48	.77	.30	.77	.59	.46	.28
0.6	30	M. Complex	High	.56	.33	.49	.88	.34	.88	.72	.54	.31
0.6	30	Equal	Low	.47	.31	.46	.63	.37	.62	.51	.47	.26
0.6	30	Equal	Moderate	.54	.33	.52	.75	.48	.75	.63	.56	.29
0.6	30	Equal	High	.62	.36	.57	.88	.64	.88	.78	.68	.33
0.6	30	M. Basic	Low	.47	.32	.45	.58	.39	.56	.50	.49	.26
0.6	30	M. Basic	Moderate	.55	.34	.52	.73	.52	.73	.63	.60	.28
0.6	30	M. Basic	High	.64	.40	.62	.87	.73	.87	.78	.75	.37
0.6	60	M. Complex	Low	.51	.31	.50	.86	.31	.86	.65	.53	.28
0.6	60	M. Complex	Moderate	.56	.32	.54	.95	.34	.95	.78	.61	.30
0.6	60	M. Complex	High	.63	.33	.53	.99	.44	.99	.89	.70	.31
0.6	60	Equal	Low	.56	.32	.52	.84	.53	.84	.69	.63	.27
0.6	60	Equal	Moderate	.62	.34	.59	.93	.67	.93	.83	.73	.31
0.6	60	Equal	High	.70	.38	.64	.98	.83	.98	.93	.83	.34
0.6	60	M. Basic	Low	.57	.33	.52	.79	.57	.78	.69	.67	.28
0.6	60	M. Basic	Moderate	.65	.36	.61	.91	.73	.91	.83	.79	.33
0.6	60	M. Basic	High	.73	.46	.71	.98	.64	.98	.94	.88	.41

"True model" indicates the estimation of data belonging to the models. The subsequent two columns indicate the retrofitting estimations of the true model data.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 97-111

Çok Kategorili Bilişsel Tanı ve Çok Boyutlu Madde Tepki Kuramı Modellerinin Karşılıklı Uyarlanması *

Levent YAKAR** Nuri DOĞAN *** Jimmy DE LA TORRE ****

Öz
Bu araştırmada çok kategorili bilişsel tanı ve çok boyutlu madde tepki kuramı (ÇBMTK) modellerinin birbiri yerine kullanımı durumunda birey parametrelerinin tekrar elde edilebilirliği incelenmiştir. Bu amaç doğrultusunda çok kategorili bilişsel tanı modellerinden polytomous generalized deterministic input noisy and gate (pG-DINA) ve fully-additive model (fA-M) ve ÇBMTK modellerinden ise telafi edici 2PL modeli için veriler üretilmiştir. Veriler, madde ayırt edicilik indeksi, maddelerin yapılarına göre testteki oranı, test uzunluğu ve yetenekler arası korelasyon değerleri farklılaştırılarak toplamda 54 koşul kullanılarak 25 tekrar ile üretilmiştir. Verinin her üç modelle de kestirimi sonucu ortaya çıkan birey parametreleri ile veri üretiminde kullanılan birey parametrelerinin karşılaştırılması ile bulgular elde edilmiştir. Bulgulara göre tüm veri türlerinin en yüksek doğrulukta kestirimi ait oldukları modeller tarafından gerçekleştirilmiştir. Uyarlama içeren analizlerde ise fA-M diğer iki model verilerini, ÇBMTK ise fA-M verisini, verinin ait olduğu model kestirimine yakın bir doğruluk oranında kestirmiştir. PG-DINA'nın diğer iki model verilerini, ÇBMTK'nin ise pG-DINA verisini kestirmede düşük performansa sahip olduğu gözlenmiştir. Araştırmada kullanılan koşullardan sırasıyla test uzunluğu ve madde ayırt ediciliğinin birey parametre doğruluğuna etki eden en kuvvetli faktörler olduğu görülmüştür. Madde yapısı oranı koşulunun ise ÇBMTK verisi analizlerinde ve uyarlamalarında çok daha etkili olduğu görülmüştür. Yetenekler arası korelasyonun varlığının birey parametre doğruluğuna etkisinin ise ÇBMTK verisinin analizlerinde daha belirgin ancak yine de sınırlı olduğu görülmüştür.

Anahtar Kelimeler: Çok kategorili bilişsel tanı modelleri, pG-DINA, fA-M, çok boyutlu madde tepki kuramı (ÇBMTK), uyarlama

GİRİŞ

Eğitimde ve psikolojide gerçekleştirilen ölçme işlemleri bir veya daha fazla özelliği ölçmeye yönelik olabilmektedir. Tek bir özelliği ölçmeye odaklanan ölçeklerin de gerçekte birden fazla özelliği ölçtüğü; ancak ölçülen özellikler arasındaki yüksek korelasyon nedeniyle ölçeğin diğer özelliklerin ölçülmesine duyarsız olabileceği ifade edilmektedir (Reckase, 2007). Bu durumda testin sadece başat faktör olarak yer alan özelliği ölçtüğü söylenebilir. Ancak özellikle psikolojik temelli ölçümlerde ölçülen özellikler arası korelasyon, her zaman başat faktörün diğerlerini baskılamaya yetecek kadar çok yüksek olmayabilir. Bu durumda farkında olarak ya da olmayarak birden fazla özelliğin ölçülmesi söz konusu olabilmektedir.

Birden fazla özelliğin ölçümü gerçekleştirilmek istendiğinde ölçüm sonucunu puanlara dönüştürecek pek çok psikometrik model bulunmaktadır. Bu psikometrik modellerden biri olan madde tepki kuramında (MTK) genellikle tek boyutluluk varsayımına sahip tek boyutlu MTK modelleri ilk olarak

* Bu çalışma ilk yazarın, diğer yazarların danışmanlığında gerçekleştirdiği "Bilişsel tanı ve çok boyutlu madde tepki kuramı modellerinin karşılıklı uyumlarının incelenmesi" başlıklı doktora tezinden üretilmiştir.

** Dr. Öğr. Üyesi, Kahramanmaraş Sütçü İmam Üniversitesi, Eğitim Fakültesi, Kahramanmaraş, Türkiye, l_yakar@hotmail.com, ORCID ID: 0000-0001-7856-6926
*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara, Türkiye, nuridogan2004@gmail.com, ORCID ID: 0000- 0001-6274-2016
**** Profesör, The University of Hong Kong, Eğitim Fakültesi, Hong Kong, Çin, j.delatorre@hku.hk, ORCID ID: 0000-0002-0893-3863

Bu makaleye atıfta bulunmak için:
Yakar, L., Doğan, N., & de la Torre, J. (2021). Retrofitting of polytomous cognitive diagnosis and multidimensional item response theory models. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 97-111. doi: 10.21031/epod.778861

Geliş Tarihi: 10.08.2020
Kabul Tarihi: 26.05.2021

akla gelmektedir. Ancak MTK ailesi içerisinde birden fazla özelliği aynı anda ölçebilecek modeller de bulunmaktadır.

Çok Boyutlu Madde Tepki Kuramı (ÇBMTK)

Ölçülen özelliğin tek boyutlu olmasını gerektiren MTK'de bu sınırlılığın giderilmesine yönelik olarak çok boyutlu testler için çok boyutlu madde tepki kuramı (ÇBMTK) geliştirilmiştir. ÇBMTK, birden fazla yapı veya boyut olarak bir arada ele alınan, gözlenemeyen değişkenler ile bireyin herhangi bir test maddesine doğru yanıt verme olasılığı arasındaki ilişki için kullanılır (Ackerman, Geirl & Walker, 2003). ÇBMTK ile her bir yeteneğe ilişkin yetenek parametreleri elde edilebilmektedir (Reckase, 2009). ÇBMTK uygulamaları çok sayıda özelliğin ölçülebilmesi nedeniyle fark yaratmış ve giderek yaygın hale gelmiştir.

ÇBMTK'de çok sayıda model bulunmaktadır. Modellere ilişkin alan yazındaki temel sınıflama, telafi edici ve telafi edici olmayan modeller olarak ön plana çıkmaktadır. Telafi edici modellerde bireylerin yüksek olan yeteneği düşük olan yeteneğinin eksiğini telafi edebilirken, telafi edici olmayan modellerde böyle bir durum söz konusu değildir. Telafi edici olmayan modelin özellikle açımlayıcı çalışmanın gerektiği durumlarda tahmin hesaplama sıkıntısı nedeniyle araştırmalarda ikinci planda kaldığı görülmektedir (Chalmers & Flora, 2014). Bu sebeple alanda telafi edici modellerin hâkimiyeti söz konusudur.

ÇBMTK modelleri arasındaki diğer bir temel farklılık olarak model analizinde kestirilecek madde parametreleri sayısı öne çıkmaktadır. Sadece madde güçlük parametresinin *(d)* kestirildiği model 1 parametreli lojistik (PL), *d* parametresinin yanı sıra madde ayırt edicilik (***a***) parametresinin de kestirildiği model 2 PL, *a* ve *d* parametreleriyle birlikte şans başarısı parametresini (*c*) de kestirildiği model ise 3 PL ÇBMTK modeli olarak adlandırılmaktadır.

McKinley ve Reckase (1982) tarafından ortaya konulan telafi edici 2PL modeli yaygın olarak kullanılan bir ÇBMTK modelidir. Telafi edici 2PL modeli için *i* bireyinin *j*. maddeyi doğru yanıtlama olasılığı aşağıdaki formülle hesaplanmaktadır.

$$p(\boldsymbol{\theta}_i, \boldsymbol{a}_j, d_j) = \frac{1}{1 + \exp(-D\sum_{k=1}^{m}(a_{jk}\theta_{ik}) + d_j)}$$

D=1,7 ölçekleme sabiti

θ_{ik}= *i* bireyinin *k*. boyuttaki yetenek parametresi

m= boyut sayısı

Bilişsel Tanı Modelleri (BTM)

MTK, ÇBMTK'den başka modellere de öncülük etmiştir. Bu modellerden biri de bilişsel tanı modelleridir (BTM). BTM'ler bilişsel tanı değerlendirme (BTD; Cognitive Diagnosis Assessment) amacını taşıyan testlere aracılık eden modeller olarak tanımlanmaktadır (de la Torre & Minchen, 2014). BTM'nin ana amacı bireylerin öğrenme eksikliklerini tespit etmek adına, bireyin testte yoklanan niteliklere sahip olup olmadığını belirlemektir. Buna göre BTM'ler genel olarak kişilerin biliş göstergelerine göre sınıflama işlemi yapmaktadır. Bu göstergelerin tek tek incelenebilmesi BTM'lerin avantajı olarak öne çıkmaktadır. BTM, MTK'ye nazaran daha küçük birimleri ayrı ayrı ölçmesi sayesinde son yıllarda giderek yaygınlaşmaktadır (Rupp, Templin ve Henson, 2010; von Davier ve Lee, 2019).

BTM'nin amacı testte yoklanmak istenen niteliklerin öğrencilerde olup olmamasına göre öğrencileri örtük kategorilerde sınıflamaktır. Bu sınıflama işlemi temelde, testi alan bireylerin yeteneklerine, maddeleri doğru cevaplamak için gerekli olan niteliklere sahip olup olmamalarına göre 0-1 değerleri atanarak gerçekleşmektedir. Bireyin maddeyi doğru yanıtlaması ilgili maddede yoklanan nitelik veya niteliklerin öğrencide bulunduğuna işaret etmektedir. Belirli bir nitelik için düşünülecek olursa bu

niteliği gerektiren maddelerin yüksek oranda doğru yanıtlanması öğrencinin niteliğe sahip olduğuna dair sonucun geçerliğini artıracaktır (Rupp, Templin ve Henson, 2010).

Çok sayıda farklı modelin bulunduğu BTM'lerin ortak yönlerinden biri Q matrisi girdisidir. Q matrisinde madde ile maddeyi yanıtlamak için bulunması gereken nitelik(ler) ilişkilendirilir. Q matrisinde satırlar maddeleri, sütunlar ise nitelikleri gösterir. Yapı olarak basit görünmesine rağmen Q matrisinin BTM'ler için oynadığı rol kritiktir. Q matrisi sadece madde nitelik ilişkisinde değil aynı zamanda bireylerin yeteneklerin belirlenmesinde önemli rol oynar. Q matrisinde temsil edilen nitelikler için bireylere profil ataması yapılır (de la Torre & Minchen, 2014).

BTM'de görülen temel sınıflamalardan birisi, Q matriste temsil edilen niteliklerin çok sayıda düzeye sahip olup olmamasına ilişkindir. Maddede yoklanan nitelik iki (0-1) veya çok (0-1-2…) kategorili yapıya sahip olabilir. İki kategorili yaklaşımda Q matristeki herhangi nitelik, ilgili maddeyi doğru yanıtlamak için gerekli veya değildir. Bu durum matriste, nitelik madde için gerekli ise 1 değilse 0 temsili ile gösterilir. Yine aynı şekilde analiz sonucu nitelik bireyde yoktur veya vardır şeklinde yapılan değerlendirme de birey profilinde 0-1 ile temsil edilir. Çok kategorili yaklaşımda ise niteliğin maddede ve bireyde bulunma seviyesi olduğu söylenebilir. Örneğin 3 kategorili bir nitelik için, bir tane sahip olmama (0) ve iki tane de farklı seviyelerde sahip olma (1,2) durumu söz konusudur. Farklı düzeylerde niteliğe sahip olma özelliği sayesinde çok kategorili nitelikler, madde güçlük parametrelerine benzemektedirler.

Çok kategorili niteliğe sahip modeller, genelde iki kategorili niteliğe sahip olan modellerin doğrudan aşamadıkları sorunları gidermek adına iki kategorili modellerin genişletilmiş halidir. Diğer bir deyişle iki kategorili modeller çok kategorili modellere dönüştürülebilmektedir. Bu modeller arasında sık kullanılan birisi de, Generalized Deterministic-Input, Noisy-And Gate (G-DINA) modelinin (de la Torre, 2011) niteliklerin ikiden fazla kategorisi olması durumunda kullanılan versiyonu olan polytomous G-DINA (pG-DINA) modelidir (Chen ve de la Torre, 2013).

Polytomous Generalized Deterministic-Input, Noisy-And Gate (pG-DINA)

BTM'ler bazı kısıtlamaların uygulanması veya kaldırılması ile birbirine dönüşebilen formüllere sahip olabilmektedirler. Yapısında kısıtlamaya sahip olmayan modeller doygun, kimi kısıtlamalara sahip olan modeller ise sınırlandırılmış model olarak adlandırılmaktadır (de la Torre ve Lee, 2013). Doygun model olarak G-DINA, bunun sınırlandırılmış formları olarak ele alınabilecek Deterministic-Input, Noisy-And Gate (DINA) (Junker ve Sijstma, 2001) ve Deterministic-Input, Noisy-Or Gate (DINO) (Templin ve Henson, 2006) bu duruma örnek olarak verilebilir. Benzer şekilde pG-DINA doygun bir model olarak önerilmiş ve farklı kısıtlamalarla, sınırlandırılmış çok kategorili modellere dönüştürülebilmektedir.

PG-DINA temelde örtük grupların her bir nitelik için sahip olduğu düzeyi, madde için gerekli düzey ile karşılaştırarak iki kategorili yapıya dönüştürür, bu işlem çökmüş nitelik vektörü olarak adlandırılır. Testte üç niteliğin yoklandığı (K=3), tüm niteliklerin üç (0,1,2) kategoriye sahip olduğu (M=3) ve birinci nitelikten 2., ikinci nitelikten 1. seviyede nitelik gerektiren, 3. niteliğin yoklanmadığı örnek madde (2,1,0) için, pG-DINA'nın parametre hesaplamalarında kullandığı çökmüş grup (Chen ve de la Torre, 2013) gösterimi Tablo 1'de verilmiştir.

Tablo 1. Orijinal Grup Vektörleri için İndirgenmiş ve Çökmüş Grup Vektörleri

Orijinal $\boldsymbol{\alpha}_{lj}$	İndirgenmiş Grup Vektörü ($\boldsymbol{\alpha}_{lj}$*)	Çökmüş Grup Vektörü ($\boldsymbol{\alpha}_{lj}$**)
(0,0,0), (0,0,1), (0,0,2)	(0,0)	(0,0)
(1,0,0), (1,0,1), (1,0,2)	(1,0)	
(0,1,0), (0,1,1), (0,1,2)	(0,1)	(0,1)
(1,1,0), (1,1,1), (1,1,2)	(1,1)	
(0,2,0), (0,2,1), (0,2,2)	(0,2)	
(1,2,0), (1,2,1), (1,2,2)	(1,2)	
(2,0,0), (2,0,1), (2,0,2)	(2,0)	(1,0)
(2,1,0), (2,1,1), (2,1,2)	(2,1)	(1,1)
(2,2,0), (2,2,1), (2,2,2)	(2,2)	

Tablo 1'de *l* örtük grubundaki bireylerin *j*. maddesi için çökmüş grupları görülmektedir. Orjinal $\boldsymbol{\alpha}_{lj}$'nin her bir elemanı, örnek maddenin nitelik vektöründe (2,1,0) karşılık gelen elemanına eşit veya büyük ise $\boldsymbol{\alpha}_{lj}$**'de 1, değilse 0 olarak yer almıştır. Maddede 3. nitelik yoklanmadığı için, madde için gerekli nitelik sayısı, *Kj** =2 olarak gerçekleşmiştir. Bu sayede $2^{Kj^*}=2^2=4$ gruptan oluşan çökmüş $\boldsymbol{\alpha}_{lj}$** elde edilmiş olur. Bu örtük sınıflar için pG-DINA formülü kullanılarak doğru yanıt verme olasılığı hesaplanır.

$$P(\boldsymbol{\alpha}_{lj}^{**}) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk}^{**} + \sum_{k'>k}^{K_j^*}\sum_{k=1}^{K_j^*} \delta_{jkk'}\alpha_{lk}^{**}\alpha_{lk'}^{**} + \cdots + \delta_{j1\cdots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lk_j^*}^{**}$$

PG-DINA Tablo 1'de verilen çökmüş grup vektörleri kullanırken, bir diğer çok kategorili BTM olan fully Additive Model (fA-M; Yakar, de la Torre ve Ma, 2017) ise indirgenmiş grup vektörlerini kullanmaktadır.

Fully Additive Model (fA-M)

Doygun yapıya sahip pG-DINA'ya kısıtlamalar uygulandığında polytomous Additive Cognitive Diagnosis Model'e (pA-CDM; Chen & de la Torre, 2013) dönüştürülebilmektedir. PG-DINA'da bulunan, bireyin maddede yoklanan niteliklerin birden fazlasına sahip olmasından kaynaklı etkileşim etkisinin kaldırılmasıyla kısıtlanmış model olan pA-CDM ortaya çıkmaktadır. PA-CDM bireyin maddede yoklanan nitelik seviyesine sahip olup olmamasına göre toplama işlemi üzerine kurulu bir yapıya sahiptir. FA-M da pA-CDM gibi eklemeli ve sınırlandırılmış bir model olarak ele alınabilir. Bu iki model arasındaki en büyük fark, madde yanıtlama fonksiyonlarını elde ettikleri örtük gruplardır. PA-CDM, pG-DINA gibi çökmüş gruplar için ayrı madde yanıtlama olasılıkları hesaplarken, fA-M ise Tablo 1'de gösterilen indirgenmiş gruplar için ayrı madde yanıtlama olasılıkları hesaplamaktadır.

Sınırlandırılmış bir model olmasına rağmen, indirgenmiş gruplar için madde yanıtlama olasılığı hesaplaması, fA-M'ı diğer pek çok BTM'lerden ayıran bir nokta olarak ön plana çıkmaktadır. FA-M'in pA-CDM'den farkının pratikteki karşılığı, bireyin niteliğe sahip olması ya hep ya hiç mantığı (0-1) yerine niteliğin seviyesine (0-1-2-…) göre farklı doğru yanıtlama olasılığı hesaplamasıdır. Modelin bu yönüyle ÇBMTK'ya yaklaştığı söylenebilir. Daha fazla bilgi, yetenek, beceriye sahip olan bireyin daha yüksek doğru yanıtlama olasılığına sahip olması ayrıca tüm MTK modelleri için de geçerlidir. FA-M'ın madde yanıtlama fonksiyonu aşağıdaki gibidir.

$$P(\boldsymbol{a}_{lj}^{**}) = \delta_{j0} + \sum_{k=1}^{K^*} \sum_{m=1}^{M_k} \delta_{jkm} \alpha_{lk}^{*}$$

δ_{j0} j maddesi için temel başarı olasılığı;
δ_{jkm} j maddesinin k. niteliğindeki m. seviye için başarı olasılığı;
K^* maddede yoklanan nitelik sayısı;
M_k k niteliğinin en yüksek seviyesi;
α_{lk}^{*} l bireyinin k niteliği j maddesinin k niteliğinden küçükse (0) değilse (1)

BTM ve ÇBMTK'nin en önemli ortak yönleri çok boyutlu ölçeklerde uygulanabilmeleridir. Bunun yanında her iki kuram da telafi edici ve telafi edici olmayan modeller barındırmaktadır (Reckase, 2009; Rupp, Templin ve Henson, 2010). Bu benzerliğin dayanak noktası modellerin birden fazla özellik veya yeteneği kestirmeleridir. Ayrıca, modellerde kullanılabilecek madde yapıları da ortaktır. Her iki model için de basit yapı ve karmaşık yapıdaki maddeler bulunmaktadır. Bu özellik modellerin analizlerinde önemli bir konuma sahiptir. BTM ve ÇBMTK'nin farklılıkları ise modellerin birbiri yerine uygulanmaları durumunda ortaya çıkabilecek olumsuzlukları meydana getirmektedir. Model farklılıklarına göz atıldığında, modellerin kullanım amaçları, madde parametreleri, birey parametrelerinin sürekli veya kesikli olma durumları ve ölçme birimlerinin büyüklükleri gibi özellikler ön plana çıkmaktadır.

Psikometrik modellerin birbirine benzerlikleri hangi modelin ne zaman kullanılacağına ilişkin soru işaretlerine yol açabilmektedir. Bu durumda birbirine yakın modellerin aynı veriye uygulanarak sonuçların incelenmesi değerlendirmeye derinlik ve zenginlik kazandıracaktır. Diğer yandan ise bu benzerlikler, testin uygulandığı modelin belli olması durumunda, teste farklı bir bakış açısı kazandırmak için testin uygulanabileceği yakın model seçeneğini sunmaktadır. Benzer altyapıya sahip modellerden elde edilen sonuçların yorumlanması, model çıktılarının karşılaştırmasıyla birlikte daha anlamlı hale gelecektir. Bu yüzden gittikçe yaygınlaşan bu iki psikometrik modele ait çıktılar karşılıklı olarak değerlendirilmesinde fayda görülmektedir.

Uyarlama (retrofit) olarak ifade edilen, verinin ait olduğu model dışında farklı bir modelle incelenmesine dayanan işlem sayesinde veriden, mevcut bilgiler dışında başka bilgiler edinilmesi amaçlanmaktadır. Bir modele aitliği bilinen veri setinin başka modele uygulaması durumunda elde edilecek yeni bilgilerle mevcut çıktıların çeşitlendirilmesi, desteklenmesi ve zenginleştirilmesi uyarlama işleminin temelini oluşturmaktadır. Uyarlama işleminde mevcut model ile uyarlama ile katkıda bulunacak modelin sonuçlar açısından birbirini destekler durumda olması işlem sonucunu değerli kılacaktır.

Alanyazın tarandığında BTM ve MTK ile ilgili pek çok uyarlama çalışmasının olduğu görülmektedir. Bunlardan bir kısmı temelde BTD amacıyla hazırlanmamış testin BTM'ye uyarlanmasını (Ardıç, 2020; Chen ve Chen, 2016; Chen ve de la Torre, 2014; Lee, Park ve Taylan 2011; Liu, Huggins-Manley, ve Bulut, 2018; Şen ve Arıcan, 2015) içermektedir. Diğer yandan BTM'lerden bir modelin MTK modelleriyle karşılıklı uyarlanmasını içeren çalışmaların (de la Torre ve Karelitz, 2009; Wang 2009) olduğu da görülmektedir. Ancak çok kategorili BTM'ler için uyarlama çalışmasına rastlanamamıştır. Ayrıca bu çalışmada iki BTM ve bir ÇBMTK modeli olmak üzere üç modelin uyarlanması da çalışmanın önemini artıran bir diğer husus olarak görülmektedir.

Birbirinden uzak modelleri birbiri yerine kullanarak uyumu yüksek sonuçlar beklemek gerçeklikten uzaktır. Daha iyi kıyaslanabilir sonuçlar için değinilen uyarlama çalışmalarında olduğu gibi farklı alandaki modellerin yakın olanlarının uygulamaya alınması doğru bir yaklaşımdır. Bu amaçla ÇBMTK modellerine yakın BTM modelleri kullanılarak geniş perspektifte uyarlama, uygulama ve değerlendirme yoluna gidilmiştir.

Araştırmanın Amacı

Bu araştırmanın amacı BTM ile ÇBMTK'yi karşılıklı uygulayarak modellerin birbirlerine uyarlanabilme düzeylerini incelemektir. Bu amaç doğrultusunda şu alt problemlere yanıt aranmıştır;

1- Farklı madde ayırt edicilik, madde yapısı, test uzunluğu ve yetenekler arası korelasyon koşullarına göre elde edilen ÇBMTK verisinin, ÇBMTK ve BTM'ler ile kestiriminde, modellerin birey parametrelerini ve yetenek düzeylerini doğru sınıflama oranı nasıldır? Modellerden elde edilen birey parametresi doğru sınıflama oranları arasında fark var mıdır?

2- Farklı madde ayırt edicilik, madde yapısı, test uzunluğu ve yetenekler arası korelasyon koşullarına göre elde edilen BTM verilerinin, ÇBMTK ve BTM'ler ile kestiriminde, modellerin birey parametrelerini ve yetenek düzeylerini doğru sınıflama oranı nasıldır? Modellerden elde edilen birey parametresi doğru sınıflama oranları arasında fark var mıdır?

YÖNTEM

Araştırmanın Türü

Görünürde herhangi bir özel uygulaması veya kullanımı bulunmayan ve öncelikle olgu ve gözlemlenebilir gerçeklerin temellerine ait yeni bilgiler edinmek için yürütülen deneysel veya teorik çalışmalar, temel araştırma olarak ifade edilmektedir (OECD, 2002). Araştırma kapsamında ele alınan modellere uygun veriler üretilip sonrasında üretilen ve uyarlanan modellerdeki performansa odaklanılmıştır. İstenilen ölçütlere uygun olarak verilerin türetilip ÇBMTK ve BTM çerçevesinde uygulamalar ve karşılaştırmalar yapılarak yeni bilgi üretilmek istendiğinden, çalışma temel araştırma niteliğindedir.

Verilerin Elde Edilmesi

Araştırmadaki verilerin üretilmesinde madde ayırt ediciliği, madde yapısı, test uzunluğu ve korelasyon koşulları maniple edilmiştir. Madde ayırt ediciliği koşulu için uniform dağılıma sahip olan düşük a ~ $U(0{,}6 - 0{,}8)$, orta a ~ $U(0{,}9 – 1{,}1)$ ve yüksek düzeyde a ~ $U(1{,}5 - 1{,}7)$ madde ayırt edicilik değerleri kullanılmıştır. Maddelerin yapılarına göre testteki oranı koşulunu oluşturan temel husus ise maddenin boyutların/niteliklerin sadece biriyle veya birden fazlasıyla ilişkili olma durumudur. Madde, tek boyutla/nitelikle ilişkili ise basit, birden fazla boyutla/nitelikle ilişkili ise karmaşık yapıdadır. Q matrisindeki maddelerin, %20'sinin (çoğunlukla karmaşık), %50'sinin (eşit) ve %80'inin (çoğunlukla basit) basit yapıda, geri kalanının ise karmaşık yapıda olmasıyla ile madde yapısı koşulu oluşturulmuştur. Test uzunluğu koşulunda ise testlerin 15 (kısa), 30 (orta) ve 60 (uzun) maddeden oluşmasına izin verilmiştir. Yetenekler arası korelasyon koşulu ise 0 (ilişki yok) ve 0,6 (orta düzeyde ilişki) olmak üzere iki faklı değer almasına izin verilmiştir. Gerçek veri ve telafi edici model kullanımında korelasyonun 0 olması mümkün olmamasına rağmen bu değer, ilişkinin olmadığı durumu yansıtması ve yetenekleri ilişkili durumundaki parametrelerin daha iyi anlaşılmasına yardımcı olduğu için seçilmiştir. Tüm koşulların seçiminde, alan yazındaki diğer benzer çalışmalarda (Chen ve de la Torre, 2013; Wang, 2009) kullanılan koşullar ve model performanslarına etki eden unsurlar göz önünde bulundurulmuştur. Çok kategorili verilerde analiz işleminin iki kategoriliye daha uzun sürmesi ve BTM araştırmalarında sıkça (de la Torre ve Douglas, 2004; de la Torre ve Douglas, 2008; Huebner ve Wang, 2011) kullanılması nedeniyle çalışma 25 tekrarla yapılmıştır. Böylece çalışmada, üç farklı modelde, madde ayırt ediciliği, madde yapısı ve test uzunluğu koşullarına göre üçer, yetenekler arası korelasyon göre iki koşul kullanılmış ve toplamda (3*3*3*3*2*25) 4050 veri üretilmiş ve her üç modelde analiz edilmiştir.

ÇBMTK verisinin elde edilmesi

Çaprazlanan her bir koşul için çok değişkenli normal dağılıma sahip birey parametresine ilişkin iki boyutlu 2PL ÇBMTK verisi R programında yazılan kodla üretilmiştir. Çok kategorili BTM'lerdeki

nitelik düzeyleri madde zorluk derecesine işaret etmektedir. Bu sayede çok kategorili BTM'lerin Q matrisi madde güçlüğü parametresi bulunan diğer modellere dönüştürülebilir. Veri üretiminde, test uzunluğu ve madde yapısı koşullarına uygun oluşturulan Q matrisinin 0,67 ile çarpım sonucundan 1,34 çıkartılarak elde edilen değer madde güçlük parametresi olarak kullanılmıştır. Buna göre Q matrisinde 0, 1, 2 ve 3 değerlerine karşılık güçlük parametresinde sırasıyla -1,34, -0,67, 0 ve 0,67 değerlerine karşılık gelmiştir. Q matrisindeki 0 değeri ölçülmeyen niteliği temsil ettiğinden güçlük parametresi -1,34 olan maddenin ilgili yetenek ayırt edicilik indeksi 0'a eşitlenerek maddede ölçülmemesi sağlanmıştır. Bu yolla BTM ve ÇBMTK'nın madde güçlüklerinin eşleştirilmesi amaçlanmıştır.

Doğru sınıflama oranlarının elde edilmesinde, sürekli yapıda olan ÇBMTK birey parametresinin BTM'lerle uyumlu hale getirilmesi için kesikli hale dönüştürülmüştür. ÇBMTK birey parametresine (-0,67, 0 ve 0,67) kesme noktalarını uygulanarak, her bir boyut için 0,1,2 ve 3 kesikli değerleri elde edilmiştir. Bu kesme noktaları sayesinde bireyler her bir boyut için yaklaşık dört eşit gruba atanmıştır. Çalışmada madde parametrelerinin kestirim doğruluğunu sağlamak adına veri setleri 5000 birey için oluşturulmuştur.

BTM verilerinin elde edilmesi

Çalışmada pG-DINA ve fA-M olmak üzere iki ayrı BTM kullanılmıştır. Modellerin farklı sayıda ve yapıda madde parametresine sahip olması nedeniyle BTM verileri için ilgili koşullar için öncelikle 100000 bireylik iki boyutlu 2PL ÇBMTK verisi üretilmiştir. Ardından bu veriden R yazılımı yardımıyla pG-DINA ve fA-M modelleri için iki nitelikli, uygun yapıda ve sayıda madde parametreleri elde edilmiştir. PG-DINA verileri GDINA (Ma & de la Torre, 2016) paketi ile, fA-M verileri ise R programında yazılan özgün kod ile üretilmiştir.

Verilerin Analizi

Üretilen verilerin ÇBMTK'deki kestirimi, mirt (Chalmers, 2012) paketi aracılığıyla gerçekleştirilmiştir. Birey parametresi kestiriminde EAP yöntemi kullanılmıştır. ÇBMTK analizinin doğru sınıflama oranının elde edilmesi için kestirilen birey parametreleri, üretilen birey parametreleri ile aynı şekilde kesikli hale getirilmiştir. ÇBMTK verisinin BTM'de kestirimi, pG-DINA için GDINA (Ma & de la Torre, 2016) paketi ile, fA-M için ise yazılan özgün kod ile gerçekleştirilmiştir.

Üretilen BTM verilerinin ÇBMTK'de kestiriminde, veri orijinalde 2 PL yapısına dayanmasına rağmen, yapısı değiştiğinden dolayı ÇBMTK analizi 2PL ve 3PL modellerinden hangisinde daha iyi uyum sağladığı kontrol edilmiştir. Veri her iki modelde de analiz edildikten sonra, model-veri uyumu indekslerinden deviance indeksleri fark R yazılımı üzerinden ANOVA analizi ile test edilmiştir. Eğer fark istatistiksel olarak anlamlı (p<,05) ise 3PL modeline ait parametreler dikkate alınmıştır. Genelde 3PL modelin veri için daha iyi uyumu sağladığı görülmüştür.

Analizlerin raporlanmasında birey parametresinin doğru sınıflaması oranı (DSO) kullanılmıştır. Bireyin her iki boyuta/niteliğe ait üretilen ve kestirilen yetenek parametresinin (vektör) eşit olması, bireyin kestirilen model tarafından doğru sınıflandığı anlamına gelmektedir. Doğru sınıflanan birey sayısının üretilen birey sayısı olan 5000'e oranı, modelin DSO'sunu vermektedir. 25 tekrar ile gerçekleştirilen çalışmanın ortalama DSO'su raporlanmıştır. Ayrıca her bir model verisinin analizinde elde edilen DSO'lar arasında anlamlı fark olup olmadığı ANOVA ile test edilmiştir. Gruplar arası varyansın eşit olmaması (p<,05) nedeniyle grupların ikili karşılaştırması Tamhane tekniğiyle gerçekleştirilmiştir.

Yetenek düzeylerinin doğru sınıflanma oranı ise her bir boyut/nitelik düzeyinin model tarafından ne derece doğru kestirildiğini yansıtmaktadır. Diğer bir deyişle modelin yetenek düzeylerine göre performansını yansıtmaktadır. Bütün bireylerin iki yetenek/nitelikten sahip olduğu yetenek/nitelik düzeylerinin DSO'larının ortalamaları raporlanmıştır.

Her bir model verisinin üretiminde, temel koşullar çaprazlanarak 54 farklı koşul kullanılmıştır. Analiz sonuçlarının sunumunda ise temel koşulların yapı olarak birbirinden bağımsız olması ve bunun sonucu

koşullar arası etkileşim olmaması nedeniyle, çaprazlanmış koşullar yerine temel koşulların farklı düzeylerindeki DSO ortalamaları kullanılmıştır. Böylece bulguların sunumunda ve okunmasında kolaylık sağlanmıştır. Çaprazlanmış 54 koşulun, 25 tekrardaki DSO ortalamaları EK 'te sunulmuştur.

BULGULAR

Bu bölümde birinci alt problemdeki BTM'lerin ÇBMTK verisine uyarlanmasını ve ikinci alt problemdeki ÇBMTK'nin BTM verilerine uyarlanmasını içeren analiz sonuçları yer almaktadır.

ÇBMTK Verisinin Analizine İlişkin Sonuçlar

ÇBMTK verisinin analizi ile elde edilen DSO'lar Tablo 2'de sunulmuştur.

Tablo 2. ÇBMTK Verisinin Analizinde Elde Edilen DSO'lar

Koşul	Düzey	ÇBMTK	pG-DINA	fA-M
Madde Ayırt Edicilik	Düşük	0,43	0,27	0,40
	Orta	0,50	0,29	0,45
	Yüksek	0,58	0,33	0,48
Madde Yapısı	Çoğunlukla Karmaşık	0,45	0,27	0,38
	Eşit	0,52	0,29	0,43
	Çoğunlukla Basit	0,54	0,33	0,51
Test Uzunluğu	15	0,41	0,28	0,36
	30	0,50	0,30	0,44
	60	0,60	0,32	0,52
Yetenekler Arası Korelasyon	0	0,48	0,26	0,38
	0,6	0,53	0,33	0,50

Tablo 2 incelendiğinde ÇBMTK'nın kendi verisinin analizinde tüm koşullarda %41-60 DSO ile en yüksek performansı gösterdiği ve bunu %36-52 DSO ile fA-M uyarlamasının izlediği görülmektedir. PG-DINA'nın ise %26-33 ile tüm koşullarda en düşük DSO'lara sahip olduğu görülmektedir. DSO'lar incelendiğinde ÇBMTK analizi ve fA-M uyarlaması oranlarının birbirine yakın, fA-M ve pG-DINA uyarlamaları arasındaki farkların ise daha fazla olduğu görülmektedir.

Madde ayırt edicilik indekslerinin artışı tüm modellerin daha iyi performans göstermelerini sağlamıştır. ÇBMTK analizinde performans artışının (,15) BTM'lere (,06-,08) göre daha fazla olduğu gözlenmektedir. Benzer şekilde, basit yapıdaki maddelerin testteki oranı arttıkça modellerin DSO'larının arttığı görülmektedir. Artış modellere göre incelendiğinde ise fA-M'ın performansındaki artışın (0,13) diğerlerinden (,09-,06) daha yüksek olduğu gözlenmektedir. Test uzunluğu koşuluna bakıldığında, test uzunluğu artışının da DSO artışına neden olduğu görülmektedir. ÇBMTK ve fA-M için bu koşul en düşük (,41-,36) ve en yüksek (,60-,52) performansı barındırmaktadır. PG-DINA uyarlamasındaki artış (,04) ise son derece sınırlı olmuştur. İncelenen diğer koşul olan yetenekler arası korelasyonun varlığında da DSO'ların arttığı görülmektedir. Artışın fA-M (,12) başta olmak üzere BTM uyarlamalarında daha belirgin olduğu gözlenmektedir.

Modellerden elde edilen birey parametresi doğru sınıflama oranları arasında farkın istatistiksel olarak anlamlılığına ilişkin test sonucu Tablo 3'te sunulmuştur.

Tablo 3. ÇBMTK Verisinin Analizi Sonucu Elde Edilen DSO'lar Arası Farkın Testi

Varyans Kaynağı	Kareler Toplamı	sd	*F*	Fark
Gruplar arası	30,807	2	1592,984*	ÇBMTK>fA-M
Grup içi	39,132	4047		ÇBMTK>pG-DINA
Toplam	*69,939*	*4049*		*f-AM>pG-DINA*

*p<,001

Tablo 3'teki değerlere göre ÇBMTK verisinin analizi sonucu elde edilen DSO'lar arasında anlamlı fark bulunmuştur [F(2,4047)=1592,984, p<,001]. Modellerden edilen DSO'lar Tamhane yöntemi ile ikili karşılaştırıldığında, ÇBMTK sonuçlarının BTM sonuçlarından, fA-M sonuçlarının ise pG-DINA sonuçlarından anlamlı olarak daha yüksek olduğu görülmüştür (p<,001).

ÇBMTK verisinin analizi sonucu yetenek düzeylerinde gerçekleşen DSO'lar Tablo 4'te sunulmuştur.

Tablo 4. ÇBMTK Verisinin Analizinde Yetenek Düzeylerine Göre Elde Edilen DSO'lar

Model	Düzey 0	Düzey 1	Düzey 2	Düzey 3
ÇBMTK	0,79	0,61	0,61	0,79
pG-DINA	0,96	0,10	0,10	0,96
fA-M	0,74	0,56	0,56	0,74

Tablo 4'teki sonuçlarda en çok dikkat çeken durum pG-DINA'nın düzey 0 ve düzey 3'deki bireyleri çok yüksek (,96) bir oranda doğru sınıflamasına rağmen düzey 1 ve düzey 2'de bulunan bireylerin sadece %10'unu doğru sınıflayabilmektedir. Diğer modellerde de düzey 0 ve 3 deki bireyler (,79-,74) düzey 1 ve 2'deki bireylerden (,61-,56) daha yüksek bir yüzdede doğru sınıflanmaktadır. Ancak bu farkın pG-DINA'daki kadar büyük olmadığı, daha dengeli oranların olduğu görülmektedir.

pG-DINA Verisinin Analizine İlişkin Sonuçlar

PG-DINA verisinin analizi ile elde edilen DSO'lar Tablo 5'te sunulmuştur.

Tablo 5. pG-DINA Verisinin Analizinde Elde Edilen DSO'lar

Koşul	Düzey	pG-DINA	ÇBMTK	fA-M
Madde Ayırt Edicilik	Düşük	0,60	0,35	0,59
	Orta	0,72	0,44	0,72
	Yüksek	0,84	0,54	0,82
Madde Yapısı	Çoğunlukla Karmaşık	0,73	0,31	0,70
	Eşit	0,72	0,50	0,71
	Çoğunlukla Basit	0,71	0,52	0,70
Test Uzunluğu	15	0,53	0,32	0,51
	30	0,73	0,45	0,72
	60	0,90	0,56	0,89
Yetenekler Arası Korelasyon	0	0,71	0,44	0,70
	0,6	0,73	0,45	0,72

Tablo 5 incelendiğinde pG-DINA'nın kendi verisinin analizinde tüm koşullarda %53-90 DSO ile en yüksek performansı gösterdiği, fA-M uyarlamasının ise %51-89 DSO ile çok yakın bir performansa sahip olduğu görülmektedir. ÇBMTK uyarlaması ise %31-56 ile tüm koşullarda en düşük DSO'lara sahip olduğu görülmektedir. DSO'lar incelendiğinde pG-DINA analizi ve fA-M uyarlaması oranlarının arasında görülen farkların (%0-2) birbirine çok yakın olduğu, fA-M ve ÇBMTK uyarlamaları arasındaki farkların ise çok büyük olduğu görülmektedir.

Madde ayırt edicilik indekslerinin artışı tüm modellerin daha iyi performans göstermelerini sağladığı görülmektedir. BTM analizlerinde performans artışlarının (,24-,23) ÇBMTK analizine (,19) göre biraz daha fazla olduğu gözlenmektedir. Basit yapıdaki maddelerin testteki oranı arttıkça sadece ÇBMTK'de DSO'larının arttığı (,21), pG-DINA analizinde çok az düştüğü (-,02), fA-M analizinde ise hemen hemen sabit kaldığı görülmektedir. Test uzunluğu koşuluna bakıldığında, test uzunluğu artışının da DSO artışına neden olduğu görülmektedir. PG-DINA ve fA-M analizleri için bu koşul en düşük (,53-,51) ve en yüksek (,90-,89) performansı barındırmaktadır. ÇBMTK uyarlamasındaki artış (,24) ise daha sınırlı olmuştur. İncelenen diğer koşul olan yetenekler arası korelasyonun varlığında da DSO'ların arttığı görülmektedir, ancak artışın çok sınırlı olduğu görülmektedir.

Modellerden elde edilen birey parametresi doğru sınıflama oranları arasında farkın istatistiksel olarak anlamlılığına ilişkin test sonucu Tablo 6'da sunulmuştur.

Tablo 6. pG-DINA Verisinin Analizi Sonucu Elde Edilen DSO'lar Arası Farkın Testi

Varyans Kaynağı	Kareler Toplamı	sd	*F*	Fark
Gruplar arası	65,273	2	1010,622*	pG-DINA> ÇBMTK
Grup içi	129,464	4009		fA-M > ÇBMTK
Toplam	194,737	4011		

*p<,001

Tablo 6'daki değerlere göre pG-DINA verisinin analizi sonucu elde edilen DSO'lar arasında anlamlı fark bulunmuştur [F(2,4009)=1010,622, p<,001]. Modellerden edilen DSO'lar Tamhane yöntemi ile ikili karşılaştırıldığında, ÇBMTK sonuçlarının BTM sonuçlarından anlamlı daha düşük olduğu görülerken (p<,001), fA-M sonuçlarının ise pG-DINA sonuçlarından farklılaşmadığı görülmüştür (p>,05).

PG-DINA verisinin analizi sonucu yetenek düzeylerinde gerçekleşen DSO'lar Tablo 7'de sunulmuştur.

Tablo 7. pG-DINA Verisinin Analizinde Yetenek Düzeylerine Göre Elde Edilen DSO'lar

Model	Düzey 0	Düzey 1	Düzey 2	Düzey 3
pG-DINA	0,87	0,81	0,80	0,87
ÇBMTK	0,65	0,56	0,60	0,71
fA-M	0,81	0,80	0,80	0,81

Tablo 7'deki sonuçlara bakıldığında pG-DINA'nın tüm düzeylerde %80-87 arasında DSO'ya sahip olduğu ve düzey 0 ve 3'deki bireyleri biraz daha yüksek oranda doğru sınıfladığı görülmektedir. FA-M ise tüm düzeylerde birbirine çok benzer şekilde %80-81 DSO'ya sahiptir. ÇBMTK analizi ise orta düzeylerde daha düşük %56-60, uç düzeylerde daha yüksek %65-71 ve kısmen de düzensiz DSO'lara sahiptir.

fA-M Verisinin Analizine İlişkin Sonuçlar

FA-M verisinin analizi ile elde edilen DSO'lar Tablo 8'de sunulmuştur.

Tablo 8 incelendiğinde fA-M'ın kendi verisinin analizinde tüm koşullarda %44-80 DSO ile en yüksek performansı gösterdiği ve bunu %39-71 DSO ile ÇBMTK uyarlamasının izlediği görülmektedir. PG-DINA'nın ise %26-34 ile tüm koşullarda en düşük DSO'lara sahip olduğu görülmektedir. DSO'lar incelendiğinde fA-M analizi ve ÇBMTK uyarlaması oranlarının birbirine yakın, ÇBMTK ve pG-DINA uyarlamaları arasındaki farkların ise daha fazla olduğu görülmektedir.

Tablo 8. fA-M Verisinin Analizinde Elde Edilen DSO'lar

Koşul	Düzey	fA-M	ÇBMTK	pG-DINA
Madde Ayırt Edicilik	Düşük	0,50	0,46	0,26
	Orta	0,61	0,55	0,29
	Yüksek	0,74	0,65	0,34
Madde Yapısı	Çoğunlukla Karmaşık	0,59	0,47	0,28
	Eşit	0,63	0,57	0,29
	Çoğunlukla Basit	0,63	0,61	0,31
Test Uzunluğu	15	0,44	0,39	0,27
	30	0,61	0,55	0,29
	60	0,80	0,71	0,32
Yetenekler Arası Korelasyon	0	0,61	0,55	0,30
	0,6	0,62	0,55	0,29

Koşulların etkisi incelendiğinde ise, madde ayırt edicilik indekslerinin artışı tüm modellerin daha iyi performans göstermelerini sağladığı görülmektedir. FA-M (,24) ve ÇBMTK (,19) analizlerinde performans artışının pG-DINA (,08) analizine göre daha fazla olduğu gözlenmektedir. Bu koşul ayrıca pG-DINA için en düşük (,26) ve en yüksek (,34) performansı barındırmaktadır. Basit yapıdaki maddelerin testteki oranı arttıkça modellerin DSO'larının arttığı görülmektedir. Artış modellere göre incelendiğinde ise ÇBMTK (,14) performansındaki artışın BTM'lere (,04-,03) göre daha yüksek olduğu gözlenmektedir. Test uzunluğu koşuluna bakıldığında, test uzunluğu artışının da DSO artışına neden olduğu görülmektedir. FA-M ve ÇBMTK analizinde bu koşula göre artışın çok büyük olduğu ayrıca bu koşulun en düşük (,44-,39) ve en yüksek (,80-,71) performansı da içerdiği görülmektedir. PG-DINA uyarlamasındaki artış (,05) ise son derece sınırlı olmuştur. İncelenen diğer koşul olan yetenekler arası korelasyonun varlığında DSO'larda dikkate değer bir değişim görülmemektedir.

Modellerden elde edilen birey parametresi doğru sınıflama oranları arasında farkın istatistiksel olarak anlamlılığına ilişkin test sonucu Tablo 9'da sunulmuştur.

Tablo 9. fA-M Verisinin Analizi Sonucu Elde Edilen DSO'lar Arası Farkın Testi

Varyans Kaynağı	Kareler Toplamı	sd	*F*	Fark
Gruplar arası	77,937	2	1934,53*	f-AM>pG-DINA
Grup içi	81,521	4047		f-AM> ÇBMTK
Toplam	159,458	4049		ÇBMTK>pG-DINA

**p*<,001

Tablo 9'daki değerlere göre fA-M verisinin analizi sonucu elde edilen DSO'lar arasında anlamlı fark bulunmuştur [$F(2,4047)=1934,53$, $p<,001$]. Modellerden edilen DSO'lar Tamhane yöntemi ile ikili karşılaştırıldığında, fA-M sonuçlarının ÇBMTK ve pG-DINA sonuçlarından, ÇBMTK sonuçlarının ise pG-DINA sonuçlarından anlamlı olarak daha yüksek olduğu görülmüştür ($p<001$).

FA-M verisinin analizi sonucu yetenek düzeylerinde gerçekleşen DSO'lar Tablo 10'da sunulmuştur.

Tablo 10. fA-M Verisinin Analizinde Yetenek Düzeylerine Göre Elde Edilen DSO'lar

Model	Düzey 0	Düzey 1	Düzey 2	Düzey 3
fA-M	0,85	0,67	0,68	0,84
ÇBMTK	0,82	0,62	0,63	0,81
pG-DINA	0,98	0,11	0,11	0,98

Tablo 10'daki sonuçlarda, ÇBMTK verisinin analizini içeren Tablo 4'teki sonuca benzer şekilde pG-DINA'nın düzey 0 ve düzey 3'deki bireyleri mükemmele yakın bir oranda (,98) doğru sınıflamasına rağmen düzey 1 ve düzey 2'de bulunan bireylerin sadece %11'ini doğru sınıflayabildiği görülmektedir. PG-DINA haricindeki modellerde de düzey 0 ve 3 deki bireyler düzey 1 ve 2'deki bireylerden daha yüksek bir yüzdede doğru sınıflanmaktadır. Ancak bu farkın pG-DINA'ya göre çok daha küçük olduğu görülmektedir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada ÇBMTK ve çok kategorili BTM'lerden pG-DINA (Chen ve de la Torre, 2013) ve fA-M'ın (Yakar, de la Torre ve Ma, 2017) birbirine uyarlanması amaçlanmıştır. Bu amaçla öncelikle madde ayırt edicilik, madde yapısı, test uzunluğu ve yetenekler arası korelasyon koşullarına göre her modele ait veriler üretilmiştir. Sonrasında her bir veri üç modelde de analiz edilerek, veri üretimindeki birey parametrelerinin analizde kullanılan model tarafından doğrulanma oranlarına bakılmıştır.

İlk alt problem için ÇBMTK'de üretilen veri beklendiği gibi en doğru ÇBMTK tarafından kestirilmiştir. Bunu fA-M kestirimi izlemiş, en düşük performans ise pG-DINA'da görülmüştür. ÇBMTK ve fA-M kestirim sonuçlarının yakın olması, ÇBMTK'de yetenek artışının maddenin doğru yanıtlama olasılığını artırmasına benzer mantığa sahip olan indirgenmiş örtük gruplarının fA-M'da kullanımıyla açıklanabilir. Çökmüş gruplar üzerinden işlem gören pG-DINA'da birey parametrelerindeki her nitelik artış madde yanıtlama olasılığına yansımayabilirken, fA-M'da her nitelik artışı maddeyi yanıtlamaya da artış olarak yansımaktadır.

Araştırmada farklılaştırılan koşullara bakıldığında ÇBMTK analizinde en yüksek verim test uzunluğu koşulundan elde edilmiştir. Bu koşulu sırasıyla madde ayırt edicilik ve madde yapısı oranı izlemiştir. Yetenekler arası korelasyon koşulunun ise diğer koşullara göre etkisi sınırlı kalmıştır. Madde yapısı koşulunun sadece ÇBTMK verisinde veya analizlerinde etkin olduğu görülmüştür. Ayrıca, BTM'lerde koşulların etki büyüklüklerinin farklılaşabileceği bulgularda ortaya çıkmıştır. PG-DINA'nın koşul değişimlerine zayıf tepkiler verdiği, fA-M'ın ise ÇBMTK'ya göre madde yapısından ve yetenekler arasında korelasyondan daha fazla etkilendiği görülmüştür. Wang'ın (2009) reduced reparameterized unified model (R-RUM; DiBello, Roussos, & Stout, 2007; Hartz, 2002) ve ÇBMTK'yı karşılıklı uyarladığı çalışmasında bulunan madde yapısı ve madde ayırt edicilik koşullarındaki değişim, fA-M uyarlamasıyla benzerlik göstermektedir. De la Torre ve Karelitz'in (2009) tek boyutlu MTK ve DINA'yı karşılıklı uyarlamasında madde ayırt edicilik koşulunda görülen değişim de yine fA-M uyarlaması ile benzerlik taşımaktadır. Bu sonuçlar birbiri ile uyumlu olan farklı modellerin karşılıklı uyarlanmasında etki eden faktörlerin ortak olabileceği çıkarımını doğurmaktadır.

İkinci alt problemde ele alınan pG-DINA verisinin analizlerine bakıldığında, pG-DINA'nın kendi verisini çok iyi bir oranda doğru kestirdiği görülmüştür. Bu oranlar çalışmada elde edilen en yüksek değerlere erişmiştir. FA-M'dan elde edilen oranların pG-DINA'ya çok yakın olduğu ve oranlar arasında istatistiksel fark olmadığı görülmüştür. PG-DINA verisinin fA-M tarafından pG-DINA benzer doğrulukta kestirilmesi, araştırmada görülen en iyi uyarlama olması bakımından dikkat çekicidir. ÇBMTK ise bu modellerden daha düşük oranlara sahiptir. Bu sonuç ilk araştırma problemindeki sonucu destekler niteliktedir. ÇBMTK ve pG-DINA'nın karşılıklı uyarlanmasında elde edilen sonuçlar istenilen düzeyde değildir. PG-DINA verisinin fA-M tarafından başarılı kestirimi, pG-DINA'da var olan etkileşim etkisinin, fA-M'da olmamasına rağmen her düzey için kestirilen etkilerce ikame edilmesi sayesinde gerçekleşmektedir. Modellerde birebir aynı madde parametreleri olmasa dahi farklı parametreler model değişiminde oluşan boşluğu doldurabilmektedir. Bu durum daha fazla sayıda madde

parametresi içeren modellerin bir avantajı olarak değerlendirilebilir. PG-DINA verisinde en düşük kestirimde bulunan ÇBMTK'nin yine de kendi verisinin analizi sonucu elde ettiği değerlere yakın düzeyde DSO'ya sahip olduğu görülmüştür. Bu açıdan bakıldığında ÇBMTK'nin pG-DINA verisinde başarısız olduğu söylenemeyebilir.

PG-DINA verisi analizinde koşullara göre ortaya çıkan farklı sonuçlara bakıldığında, test uzunluğu ve madde ayırt edicilik artışının tüm modellerin DSO'larına iyi yansıdığı görülmektedir. Buna ek olarak madde yapısındaki basitleşmenin ise sadece ÇBMTK'de artışı beraberinde getirmektedir. Burada dikkat çeken diğer bir sonuç ise, madde yapısı oranı koşulunun pG-DINA verisinde diğer model verilerine kıyasla etkisini yitirmesidir.

FA-M verisinin analizlerinde elde edilen sonuçlara bakıldığında diğer modellerde olduğu gibi yine verinin ait olduğu modelde analizi en iyi sınıflama oranına sahiptir. ÇBMTK ise az bir farkla fA-M sonuçlarını izlemektedir. Bu sonuç ÇBMTK verisi analizinde ortaya çıkan duruma benzerlik göstermektedir. Ortaya çıkan sonuçlar gerekli durumlarda ÇBMTK ve fA-M'ların birbiri yerine kullanılabileceğini göstermiştir. pG-DINA sonuçları ise ÇBMTK verisi analizinde olduğu gibi yine en düşük oranlara sahip olduğu görülmüştür. Kullanılan koşullar açısından bakıldığında ise boyutlar arasındaki ilişki haricindeki koşulların model analizlerinde etkin olduğu görülmüştür.

PG-DINA'nın gösterdiği düşük performansın orta düzeydeki bireylerde görülen çok düşük DSO'lara bağlı olduğu, yetenek düzeylerine göre sınıflama oranlarında ortaya çıkmıştır. Yapılan ekstra bir incelemede, pG-DINA kendi verisi haricindeki diğer verilerde kestiriminde orta düzeydeki bireyleri çok büyük oranda uç düzeylerde kestirdiği görülmüştür (bulgularda sunulmamıştır). PG-DINA'nın kendi verisi dışındaki veri analizlerinde orta düzeydeki bireyleri kestirimde son derece başarısız olduğu, uç düzeylerdeki başarısını ise tüm grubu merkezden uzaklaştırarak uç düzeylere yığmasına sayesinde gerçekleştirdiği söylenebilir.

Çalışmada normal dağılımdaki yetenekler uniform dağılım oluşturacak şekilde kesme puanları ile kesikli hale getirilmiştir. Buna göre tüm düzeylerde yaklaşık eşit sayıda birey bulunmaktadır. Ancak uygulamada karşılaşılması kuvvetle muhtemel kategorik yeteneklerin de normal dağılıma yakın olması halinde pG-DINA'nın uyarlamalarda çok daha düşük DSO'lara sahip olacağı ön görülmektedir. Bu yüzden pG-DINA'nın kendi modeline ait olmayan gerçek veride kullanılması daha düşük sonuçlara neden olabilir. Bu nedenle pG-DINA'nın diğer modellere ait verilerde kullanımına dikkat edilmelidir.

Verilerin üretildikleri model ile analiz edilmesi durumunda, ÇBMTK'nin, pG-DINA ve fA-M'a göre daha düşük doğrulama oranına sahip olduğu görülmüştür. Bu sonuç ÇBMTK verisi kestiriminin diğer model verilerine göre daha zor olduğu sonucunu doğurmaktadır. Bu durum ÇBMTK'nin birey parametrelerinin sürekli yapıda olduğu ancak sonuçların karşılaştırabilmek adına kesikli hale getirilmesinden kaynaklanabilir. Bu dönüşüm nedeniyle verideki bilgi ve kalite kaybı analizlere olumsuz yansımış olabilir.

Modellerin en iyi performanslarını kendi verisinde göstermesi beklenmektedir. Ancak fA-M analizlerinin pG-DINA verisinde elde ettiği DSO'lar beklenmeyen bir şekilde fA-M verisinde elde ettiğinden daha yüksek olduğu görülmüştür. Benzer bir durum ÇBMTK için geçerlidir. ÇBMTK analizinin, ÇBMTK verisindeki performansı fA-M verisinde gösterdiğinden daha düşüktür. Benzer bir sonucun Wang (2009) çalışmasındaki R-RUM ve ÇBMTK uyarlamasında ve de la Torre ve Karelitz'in (2009) tek boyutlu MTK ve DINA uyarlamasında da olduğu görülmektedir. Bu durumun sebebinin modellerin kendi verisinden daha basit verilerde daha iyi kestirim yapması olduğu düşünülmektedir. Bu genellemenin gerçekleşmesi için model-veri uyumunun da sağlanması gerekmektedir. Bu koşulun yerine gelmemesi nedeniyle ÇBMTK pG-DINA verisinde düşük bir kestirime sahiptir. Bu bilgilere göre genel yapıları açısından modeller, basitten karmaşığa göre pG-DINA, fA-M ve ÇBMTK olarak sıralanabilir.

Sonuçlar özetlenecek olursa verinin ait olduğu modelin veriyi analizi, diğer modellerin analizinden başarılı olmuştur. FA-M ise kendine ait olmayan her iki veride de verinin ait olduğu modele yakın performans sergilemiştir. Bu durum fA-M'ın pG-DINA ve ÇBMTK'ya başarılı bir şekilde uyarlanabildiğini ve gerekli durumlarda bu modeller yerine kullanılabileceğini göstermiştir. PG-DINA

ve ÇBMTK karşılıklı uyarlama çalışmalarının her ikisi de düşük doğrulama oranlarına sahiptir. Bu yüzden bu iki modelin birbiri yerine kullanımının sorunlu olduğu görülmüştür.

Bu çalışmanın sınırlılığı açısından bakıldığında ÇBMTK verisinden çok düzeyli bilişsel tanı bilgisi alınmak istendiğinde fA-M'ın iyi sonuç verdiği görülmüştür. Ancak, bu sonuç bu uyarlama işlemi yalnızca yeteneklerin çok düzeyli niteliklere dönüştürülebileceği durumlarda geçerlidir. Çok düzeyli niteliklerin ÇBMTK'ye uyarlamasında ise fA-M verileri ÇBMTK tarafından yüksek doğrulukla analiz edilmiştir. Bu uyarlama işlemi ise sadece, niteliklerin bilinen bir şekilde boyut/boyutlarda yer alabilmesi durumunda uygulanabilir. Bu şartların oluşmadığı durumlarda farklı modellerle uyarlama işlemi yapılmalıdır.

KAYNAKÇA

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-51.

Ardıç, E. Ö. (2020). *Bilişsel tanı ve çok boyutlu madde tepki modellerinin sınıflama doğruluğu ve parametrelerinin karşılaştırılması [Comparison of classification accuracy and parameters of cognitive diagnostic and multidimensional item response models]*. Yayınlanmamış Doktora Tezi, Hacettepe Üniversitesi, Ankara.

Chalmers, R. P., (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chalmers, R. P., & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement, 38*(5), 339-358.

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly, 13*(3), 218-230.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419-437.

Chen, J., & de la Torre, J. (2014). A procedure for diagnostically modeling extant large-scale assessment data: The case of the programme for international student assessment in reading. *Psychology, 5*(18), 1967-1978.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595-624.

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46*(4), 450-469.

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*(4), 355-373.

de la Torre, J., & Minchen N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*(2), 89-97.

DiBello, L.V. Roussos L. A., & Stout, W. (2007). *Review of cognitively diagnostic assessment and a summary of psychometric models*. Rao, C. Sinharay, S. (Eds.) Handbook of Statistics, Psychometrics. Vol. 26. North-Holland: Amsterdam.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*, Yayınlanmamış Doktora Tezi, University of Illinois at Urbana-Champaign.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407-419.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*(2), 144-177.

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*(3), 357-383.

Ma, W. & de la Torre, J. (2016). GDINA: The generalized DINA model framework. R package version 0.9.2.

McKinley R. L. & Reckase M. D. (1982) *The use of the general Rasch model with multidimensional item response data* (Research Report: ONR 82-1). American College Testing, Iowa City, IA.

Organisation for Economic Co-operation and Development. (2002). Frascati Kılavuzu. Paris: OECD.

Reckase, M. D. (2007). Multidimensional item response theory. Rao, C. Sinharay, S. (Ed.) *Handbook of Statistics*, Psychometrics. Vol. 26. North-Holland: Amsterdam.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

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

Şen, S., & Arıcan, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology, 6*(2), 238-253.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

von Davier, M., & Lee, Y. S. (2019). Introduction: From latent classes to cognitive diagnostic models. In *Handbook of Diagnostic Classification Models* (pp. 1-17). Springer, Cham.

Wang, Y. C. (2009). *Factor analytic models and cognitive diagnostic models: How comparable are they? – A Comparison of R-RUM and compensatory MIRT model with respect to cognitive feedback.* Yayınlanmamış Doktora Tezi, The Faculty of The Graduate School at The University of North Carolina at Greensboro).

Yakar, L., de la Torre, J., & Ma, W. (2017). *An empirical comparison of two cognitive diagnosis models for polytomous attributes*. Annual Meeting of National Council on Measurement in Education Kongresinde sunulan bildiri. National Council on Measurement in Education (NCME), San Antonio, TX.

Ek. Çaprazlanmış Koşulların Ortalama DSO'ları

Koşullar				Doğru Model	Uyarlanan Model		Doğru Model	Uyarlanan Modeller		Doğru Model	Uyarlanan Modeller	
Kore-lasyon	Test Uz.	Madde Yapısı	Madde Ayırt E.	ÇB MTK	pG-DINA	fA-M	pG-DINA	ÇB MTK	fA-M	fA-M	ÇB MTK	pG-DINA
0	15	Ç. Karmaşık	Düşük	,26	,19	,24	,39	,21	,37	,32	,28	,23
0	15	Ç. Karmaşık	Orta	,31	,20	,26	,51	,26	,48	,40	,33	,26
0	15	Ç. Karmaşık	Yüksek	,39	,21	,26	,63	,29	,60	,49	,39	,30
0	15	Eşit	Düşük	,31	,21	,27	,38	,24	,37	,34	,31	,23
0	15	Eşit	Orta	,38	,23	,29	,51	,31	,49	,43	,39	,26
0	15	Eşit	Yüksek	,47	,25	,27	,66	,43	,63	,55	,49	,30
0	15	Ç. Basit	Düşük	,32	,23	,31	,39	,26	,38	,35	,33	,23
0	15	Ç. Basit	Orta	,41	,27	,37	,51	,37	,50	,44	,43	,28
0	15	Ç. Basit	Yüksek	,52	,33	,42	,66	,53	,65	,57	,55	,35
0	30	Ç. Karmaşık	Düşük	,34	,22	,29	,58	,27	,57	,45	,40	,25
0	30	Ç. Karmaşık	Orta	,40	,22	,31	,72	,31	,71	,57	,46	,28
0	30	Ç. Karmaşık	Yüksek	,47	,23	,28	,85	,36	,83	,69	,54	,32
0	30	Eşit	Düşük	,40	,24	,35	,58	,36	,56	,50	,47	,26
0	30	Eşit	Orta	,49	,25	,38	,72	,47	,71	,62	,56	,28
0	30	Eşit	Yüksek	,59	,28	,33	,86	,63	,85	,76	,68	,33
0	30	Ç. Basit	Düşük	,43	,25	,40	,56	,38	,54	,49	,49	,26
0	30	Ç. Basit	Orta	,52	,28	,47	,71	,51	,71	,62	,60	,28
0	30	Ç. Basit	Yüksek	,63	,33	,59	,86	,73	,86	,77	,75	,37
0	60	Ç. Karmaşık	Düşük	,43	,23	,36	,81	,32	,80	,63	,53	,27
0	60	Ç. Karmaşık	Orta	,50	,23	,37	,92	,35	,91	,76	,61	,31
0	60	Ç. Karmaşık	Yüksek	,59	,24	,35	,98	,45	,97	,88	,71	,34
0	60	Eşit	Düşük	,52	,25	,43	,80	,52	,80	,68	,64	,28
0	60	Eşit	Orta	,60	,27	,48	,91	,67	,91	,81	,74	,31
0	60	Eşit	Yüksek	,69	,30	,40	,97	,82	,97	,92	,84	,37
0	60	Ç. Basit	Düşük	,55	,28	,48	,77	,56	,76	,68	,66	,29
0	60	Ç. Basit	Orta	,64	,31	,58	,90	,73	,89	,82	,79	,33
0	60	Ç. Basit	Yüksek	,72	,43	,69	,97	,60	,97	,93	,89	,42
0,6	15	Ç. Karmaşık	Düşük	,39	,29	,38	,43	,23	,42	,33	,29	,23
0,6	15	Ç. Karmaşık	Orta	,43	,31	,41	,55	,25	,54	,41	,33	,26
0,6	15	Ç. Karmaşık	Yüksek	,48	,33	,44	,66	,29	,66	,51	,39	,29
0,6	15	Eşit	Düşük	,39	,30	,38	,41	,25	,41	,35	,32	,24
0,6	15	Eşit	Orta	,45	,31	,43	,53	,31	,52	,44	,40	,26
0,6	15	Eşit	Yüksek	,52	,34	,48	,68	,43	,66	,57	,49	,30
0,6	15	Ç. Basit	Düşük	,39	,30	,38	,40	,26	,39	,35	,33	,24
0,6	15	Ç. Basit	Orta	,45	,33	,44	,52	,37	,52	,45	,43	,28
0,6	15	Ç. Basit	Yüksek	,55	,37	,53	,68	,54	,67	,58	,56	,34
0,6	30	Ç. Karmaşık	Düşük	,45	,30	,44	,65	,27	,64	,47	,40	,26
0,6	30	Ç. Karmaşık	Orta	,49	,31	,48	,77	,30	,77	,59	,46	,28
0,6	30	Ç. Karmaşık	Yüksek	,56	,33	,49	,88	,34	,88	,72	,54	,31
0,6	30	Eşit	Düşük	,47	,31	,46	,63	,37	,62	,51	,47	,26
0,6	30	Eşit	Orta	,54	,33	,52	,75	,48	,75	,63	,56	,29
0,6	30	Eşit	Yüksek	,62	,36	,57	,88	,64	,88	,78	,68	,33
0,6	30	Ç. Basit	Düşük	,47	,32	,45	,58	,39	,56	,50	,49	,26
0,6	30	Ç. Basit	Orta	,55	,34	,52	,73	,52	,73	,63	,60	,28
0,6	30	Ç. Basit	Yüksek	,64	,40	,62	,87	,73	,87	,78	,75	,37
0,6	60	Ç. Karmaşık	Düşük	,51	,31	,50	,86	,31	,86	,65	,53	,28
0,6	60	Ç. Karmaşık	Orta	,56	,32	,54	,95	,34	,95	,78	,61	,30
0,6	60	Ç. Karmaşık	Yüksek	,63	,33	,53	,99	,44	,99	,89	,70	,31
0,6	60	Eşit	Düşük	,56	,32	,52	,84	,53	,84	,69	,63	,27
0,6	60	Eşit	Orta	,62	,34	,59	,93	,67	,93	,83	,73	,31
0,6	60	Eşit	Yüksek	,70	,38	,64	,98	,83	,98	,93	,83	,34
0,6	60	Ç. Basit	Düşük	,57	,33	,52	,79	,57	,78	,69	,67	,28
0,6	60	Ç. Basit	Orta	,65	,36	,61	,91	,73	,91	,83	,79	,33
0,6	60	Ç. Basit	Yüksek	,73	,46	,71	,98	,64	,98	,94	,88	,41

Doğru model, verinin, üretildiği model ile analizini, sağındaki ilk iki sütundaki uyarlanan modeller ise doğru modele ait verinin diğer modellere uyarlamasını içermektedir.