Pegem Journal of Education and Instruction, Vol. 14, No. 3 , 2024 (pp. 103-116)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Educational Implications of Comparing Unidimensional and Multidimensional Item Response Theories

Şeyma Erbay Mermer[1]

Bilecik Şeyh Edebali Univesity, Faculty of Health Sciences, Bilecik, Turkey

ABSTRACT

This study aims to compare item and student parameters of dichotomously scored multidimensional constructs estimated based on unidimensional and multidimensional Item Response Theory (IRT) under different conditions of sample size, interdimensional correlation and number of dimensions. This research, conducted with simulations, is of a basic nature. The standard errors of the item and student parameters estimated according to both models were evaluated by the mean of the error squares. The results of the study indicate that there is no significant difference between the errors of the item parameters obtained from the unidimensional and multidimensional IRT in the cases of two-dimensional data structures and high interdimensional correlations. However, when the data structure are three and five dimensional, the item parameter errors resulting from the unidimensional IRT increase. Standard errors for item parameters decreased as the sample size increased. When the standard errors arising from ability parameters were analysIRTed, it was observed that multidimensional IRT estimated with lower errors for all conditions. As a result of the study, it is established that multidimensional IRT provides more accurate results in the analysis of multidimensional constructs, especially in the estimation of student parameters and in making decisions about students. Therefore, it is recommended that multidimensional models should be used for student ability estimation in national and international large-scale examinations.

Keywords: Multidimensional IRT, unidimensional IRT, multidimensional data, dimensionality

INTRODUCTION

The main purpose of measurement studies in education and psychology is to determine the extent to which students have latent variables. There is no single approach accepted by everyone in determining latent variables (Crocker & Algina, 1986). The 1905 Classical Test Theory (CTT) based on Charles Edward Spearman, the Item Response Theory (IRT), which is an extension of CTT and emerged in the 1930s, and the Generalizability Theory based on variance analysis are the main approaches. Since the differences in the ability levels of students affect their test performances, it is not possible to mention the invariance of the parameters estimated on the basis of the CTT (Gül, 2015). The fact that item statistics are not dependent on the group, that chance success is taken into account, and that different error estimates are made for each ability range makes ITC more advantageous compared to others (Hambleton & Swaminathan Rogers, 1991).

IRT explains the relationship between an student's responses to an item and their ability level regarding the trait to be measured with a mathematical function. In IRT, θ is called "ability/knowledge" and takes a value between -∞ and +∞ and allows for sample-independent measurements given that the assumptions are met (de Ayala, 2009). Appropriate dimensionality, local independence and monotonicity assumptions are included in the IRT.

Dimensionality is a situation related to the structure of the feature to be measured in a measurement process. While one dimension refers to a single feature, multidimensionality means measuring more than one feature (Hasançebi, Terzi, & Küçük, 2020). The assumption of unidimensionality in psychological constructs is a difficult assumption to meet.

Local independence, another assumption of the IRT, means that the student responses the items independently of each other. This assumption is tested by taking a certain ability level as a constant and looking at the correlation of the scores obtained from the responses of students at that ability level (Crocker & Algina, 1986; Embretson & Reise, 2000; Lord & Novick, 1968; McDonald, 1999).

Corresponding Author e-mail: sey.erbay@gmail.com

https://orid.org/0000-0002-7747-9545

How to cite this article: Erbay Mermer, Ş (2024). Educational Implications of Comparing Unidimensional and Multidimensional Item Response Theories. Pegem Journal of Education and Instruction, Vol. 14, No. 3, 2024, 103-116.

Source of support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest: None

DOI: 10.47750/pegegog.14.03.10

Received: 09.11.2023

Accepted: 05.01.2023 Publised : 01.07.2024

In order to provide the advantages mentioned in IRT, first of all, the most appropriate IRT model should be selected for the test data (Hambleton & Swaminathan, 1985). When the appropriate model is not selected for the test data, item parameters will be incorrectly estimated.

Recent research and literature demonstrate that multidimensional IRT is increasingly accepted in the fields of education and psychology (Hori, Fukuhara & Yamada, 2020). In this context, rather than assuming that tests or test items measure only one latent trait and making decisions accordingly, this assumption has started to be questioned. While unidimensional IRT has difficulty in meeting this challenging assumption, it has been determined that more precise, valid and reliable results are obtained with the use of multidimensional IRT models (Köse, 2012). However, the fact that the software that graphically indicates the probability of the respondent's correct answer to the item in multidimensional IRT is still under development and cannot fully serve the researchers continues to be an important shortcoming of the studies in this field.

The main purpose of this study is to determine the most appropriate model for the data set by comparing unidimensional IRT and multidimensional IRT models under the specified conditions. In particular, the extent to which the dimension difference affects the item predictions calculated within the scope of the study and possible error cases have been examined in detail. These analyses are conducted to evaluate the performance of both models and to determine the most appropriate one for the structural characteristics of the data set. In this framework, the effect of models with different dimensions on the data set and how this effect is reflected on item estimates are discussed in detail.

METHOD

Research Design

The present research is a simulation study comparing the performance of unidimensional IRT and multidimensional IRT methods under various conditions. In studies conducted in the fields of education and psychology, it is established that analyzing with the right dimension affects item predictions. For this present study, unidimensional IRT and multidimensional ICA data fits of a scale organized with simulated data according to the 2PL model were examined, and the item and test information functions of the estimation of the parameters of the appropriately estimated model were examined.

Simulation Conditions

The data set used in the research consists of simultaneous data generated with the WinGen program. In this context, according to the 2PL model, a data set consisting of 20 items with a 1-0 item-score matrix and 520 respondents is generated through the program.

Data Analysis

With the data of the current study, it is first examined whether the assumptions of unidimensional IRT and multidimensional IRT are met. Model comparisons ae made within the framework of the estimations obtained. FACTOR, IRTPRO 2.1 and R software are utilized in the calculations of the present study.

While unidimensionality, local independence and invariance assumptions were considered for unidimensional IRT, local independence and monotonicity assumptions are taken into account for multidimensional IRT. Furthermore, tetrachoric correlation values, fit indices, skewness and kurtosis values, inter-item correlation, parameter estimates, item information functions and test characteristic curves are examined for both applications.

In order to apply multidimensional IRT to the existing data set, some assumptions must be met in addition to the multidimensionality of the data set. One of these assumptions is local independence, which has the same meaning as in the unidimensional IRT and is not elaborated again here. Another assumption is monotonicity. The monotonicity assumption is that the probability of an student giving a correct answer increases with the increase of any element in the θ-vector (Reckase, 2009). That is, as any ability associated with the item increases, the probability of the student answering this item correctly also increases. Therefore, there should be a monotonic increasing relationship between the probability of a correct answer and θ. The monotonicity assumption can be satisfied in unidimensional IRT, but there are also unidimensional IRT models that do not require this assumption. Therefore, monotonicity is not considered as an assumption in unidimensional IRT (Roberts et al., 2000; as cited in: Reckase, 2009).

FINDINGS

First off, the reliability coefficients are examined for the UDIRT (Unidimensional Item Response Theory) and the MDIRT (Multdimensional Item Response Theory) . This value is calculated as .947 for unidimensional IRT and .934 for multidimensional IRT. It can be asserted that these values are quite good values for the achievement test.

In this study, 0.30 is taken as the limit value for factor loadings (Brown, 2015). In the analysis, the KMO value is calculated .949 and according to this value, it is seen that the data set has a large enough sample for factor analysis (p=0.00). In Bartlett's test of sphericity, the value is calculated

6247.2 (p=.000) and is significant at .01 level. This means that the data are normally distributed.

Skewness and Kurtosis coefficients are examined in unidimensional and multidimensional factors and it is observed that the measurement values obtained by ranking or classification for both dimensions are asymmetric and skewed and these values are larger than -1 and 1 values. Therefore, it is deemed more appropriate to calculate tetrachoric correlation coefficients when looking at correlations (Muthen, B. & Kaplan D., 1985).

Unidmensionality

This assumption has been tested for unidimensional IRT, and IRT models assuming a single latent trait are called unidimensional. Satisfying the unidimensionality assumption requires that there is a dominant factor or dimension affecting test performance. This dominant factor is assumed to be the ability to be measured by the test (Hambleton & Swaminathan, 1985). The dimensions of the items scored 1-0 in the test are tried to be determined using the FACTOR program.

Table1: Factor loadings and explained variance ratios for unidimensional IRT.

Variable	*Eigenvalues*	*Cumulative Proportion of Variance*
1	9,418	.464
2	1,023	
3	,966	
4	,851	
5	,817	
6	,762	
7	,653	
8	,532	
9	,474	
10	,452	
11	,421	
12	,386	
13	,382	
14	,353	
15	,316	
16	,283	
17	,255	
18	,252	
19	,203	
20	,155	

Goodness of Fit Index (GFI) value is examined for model and data fit index. This value is calculated .975. This value being close to 1 indicates a good fit. When we look at the calculated value, we c an interpret that our data fit the model well (Table 1).

The factor loading value for the items in the 20-item form of the test is accepted 0.30. The eigenvalue of the first of the two factors obtained is calculated 9.418 and the second 1.023. This factor explains 46.4% of the variance. Although the study seems to be two-dimensional, it can be assumed that the data set is unidimensional by observing that the eigenvalue and explained variance ratio of the dominant factor are high.

Considering that the study is two-dimensional, the KMO value is calculated .954 in the analysis and according to this value, it is seen that the data set has a sufficiently large enough sample for factor analysis (p=0.00). In Bartlett's test of sphericity, the value is calculated 6247.2 (p=.000) and is significant at .01 level. This means that the data are normally distributed.

Goodness of Fit Index (GFI) value is examined for model and data fit index. This value is calculated .989. This value being close to 1 indicates a good fit. When we look at the calculated value, we can interpret that our data fit the two-dimensional model better (Table 2).

Table 2.: Factor Loadings and Explained Variance Ratios for Multidimensional IRT.

Variable	*Eigenvalue*	*Proportion of Variance*	*Cumulative Proportion of Variance*
1	9.28745	0.46437	0.46437
2	1.71074	0.08554	0.54991
3	1.19266	0.05963	
4	1.03862	0.05193	
5	0.85432	0.04272	
6	0.72415	0.03621	
7	0.64672	0.03234	
8	0.53642	0.02682	
9	0.47542	0.02377	
10	0.44928	0.02246	
11	0.41488	0.02074	
12	0.40750	0.02038	
13	0.37212	0.01861	
14	0.34968	0.01748	
15	0.34429	0.01721	
16	0.28587	0.01429	
17	0.26771	0.01339	

Variable	*Eigenvalue*	*Proportion of Variance*	*Cumulative Proportion of Variance*
18	0.24566	0.01228	
19	0.22422	0.01121	
20	0.17231	0.00862	

The eigenvalue of the first of the two factors obtained is calculated 9.287 and the second 1.710. This factor explains 54.9% of the variance. Although the study seems to be two-dimensional, it can be assumed that the data set is unidimensional by observing that the eigenvalue and explained variance ratio of the dominant factor are high (Table 3).

Table 3: Varimax Rotated Factor Loadings for Multidimensional IRT.

Variable	*F 1*	*F 2*
V 1	0.617	0.399
V 2	0.667	0.445
V 3	0.632	0.417
V 4		0.676
V 5	0.429	
V 7	0.398	0.583
V 8		0.368
V 7	0.398	0.583
V 8		0.368
V 4		0.676
V 5	0.429	
V 6	0.734	
V 7	0.398	0.583
V 8		0.368
V 9	0.338	
V 10	0.301	0.710
V 11	0.316	0.769
V 12	0.600	
V 13	0.312	0.608
V 14	0.678	
V 15	0.437	0.620
V 16	0.658	
V 17	0.314	0.753
V 18	0.419	0.574
V 19		0.697
V 20		0.796

The factor loading value for the items in the 20-item form of the te st is accepted .30. Gaps mean values below .30. In general, it can be said that the item factor loadings after varimax rotation are at an acceptable level.

Local Independence

Local independence means that a respondent's probability of answering an item correctly is not affected by their responses to other items in the test.

For local independence, Marginal fit (X^2) and Standardized LD X^2 table are examined. The table related to this assumption is provided in Table 4).In the table

Table 4: Marginal fit (X2) and Standardized LD X2 table for Unidimensional IRT.

		Marginal										
Item	*Label*	*X2*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	S1	03										
2	S2	03	31.5									
3	S3	03	28.4	38.8								
4	S4	01	-0.2	-0.3	-0.3							
5	S5	00	7.3	6.0	2.6	3.6						
6	S6	01	2.0	2.2	2.4	3.4	-0.6					
7	S7	07	0.2	0.2	0.4	0.5	0.6	0.3				
8	S8	00	-0.4	0.2	0.3	8.7	-0.7	-0.5	3.8			
9	S9	00	-0.5	-0.5	-0.2	2.5	5.5	-0.5	1.0	-0.7		
10	S10	08	0.6	0.0	1.8	6.7	0.3	7.1	0.7	0.7	-0.1	
11	S11	08	4.2	2.8	3.3	5.3	0.4	1.3	1.7	1.5	0.9	8.3
12	S12	00	-0.4	0.0	-0.4	0.4	-0.6	5.1	0.2	-0.2	-0.7	0.4

		Marginal										
Item	*Label*	*X^2*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
13	S13	05	0.6	0.7	-0.1	3.1	1.2	5.6	2.8	0.1	2.7	0.5
14	S14	00	-0.4	-0.2	-0.2	4.3	-0.6	7.6	-0.0	0.3	2.6	0.4
15	S15	08	0.4	0.1	1.1	-0.1	0.7	0.7	5.0	0.6	0.2	0.8
16	S16	00	-0.4	-0.4	-0.3	2.3	0.9	3.9	1.2	-0.3	1.1	0.1
17	S17	11	4.9	3.8	5.1	4.6	0.2	1.7	0.5	0.3	0.6	1.5
18	S18	05	-0.3	-0.2	-0.2	1.0	0.2	-0.1	-0.1	-0.2	-0.1	-0.1
19	S19	08	3.3	0.4	2.2	2.0	-0.1	5.7	2.6	0.2	-0.1	1.2
20	S20	11	3.0	7.7	12.1	7.2	0.4	4.9	2.8	1.7	0.5	4.8

		Marginal									
Item	*Label*	*X2*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
11	S11	08									
12	S12	00	9.0								
13	S13	05	14.6	14.7							
14	S14	00	19.8	8.6	21.5						
15	S15	08	5.1	15.9	4.9	5.9					
16	S16	00	9.9	13.0	18.3	17.3	1 4.0				
17	S17	11	16.4	5.0	16.6	8.2	6.0	14.8			
18	S18	05	10.2	16.5	6.8	24.4	6.3	19.9	19.9		
19	S19	08	14.0	13.4	16.8	17.5	6.2	5.3	16.7	22.1	
20	S20	11	20.1	8.1	5.7	14.3	6.1	7.6	6.3	15.9	13.7

In the table expressing the IRTPRO standardized LD X^2 X^2 magnitude, if the correlation between two items is greater than 10, it means that the two items are dependent. In local independence, these values are expected to be less than 10. In the unidimensional IRT model, 31 values greater than 10 are found (Table 5).

Table 5: Marginal fit (X2) and Standardized LD X2 table for Multidimensional IRT Model.

		Marginal										
Item	*Label*	*X^2*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	S1	06										
2	S2	08	1.9									
3	S3	07	0.4	2.4								
4	S4	01	-0.2	0.0	-0.2							
5	S5	00	0.6	0.7	-0.2	2.8						
6	S6	00	0.0	0.8	1.2	4.0	-0.3					
7	S7	06	2.9	4.7	2.9	-0.2	3.3	0.5				
8	S8	00	-0.1	0.8	0.5	13.0	-0.6	-0.7	2.5			
9	S9	00	-0.1	0.6	0.7	3.5	4.5	-0.7	2.9	-0.7		
10	S10	10	0.2	1.8	0.2	2.0	3.7	0.8	0.9	0.1	1.0	
11	S11	10	0.2	0.1	0.4	7.7	0.0	0.3	1.3	0.2	0.0	2.5
12	S12	00	1.4	5.0	4.0	-0.4	0.2	4.8	1.0	-0.7	-0.5	2.4

		Marginal										
Item	*Label*	*X^2*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
13	S13	06	0.1	6.4	2.3	0.7	5.4	-0.1	2.1	-0.2	4.1	0.9
14	S14	01	5.1	3.5	2.9	1.1	0.0	3.7	3.2	-0.6	2.1	1.9
15	S15	06	3.1	5.1	4.8	-0.2	3.6	0.3	13.7	0.3	1.3	0.3
16	S16	01	4.0	4.3	1.9	0.1	-0.2	2.6	0.5	-0.3	1.6	1.4
17	S17	13	0.5	0.3	1.1	2.0	0.2	0.2	1.0	0.5	1.3	0.3
18	S18	04	1.5	3.0	0.5	-0.4	-0.4	1.0	-0.2	-0.1	0.4	0.0
19	S19	14	0.9	0.8	0.6	1.1	1.6	0.6	3.5	0.4	1.6	2.3
20	S20	18	1.1	1.1	4.5	2.6	0.6	2.1	3.4	0.7	0.6	2.5

		Marginal									
Item	*Label*	*X2*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
11	S11	06									
12	S12	00	0.1								
13	S13	06	1.2	-0.2							
14	S14	00	0.2	7.8	-0.2						
15	S15	10	0.3	0.1	0.4	0.1					
16	S16	00	-0.2	8.2	0.4	11.2	0.1				
17	S17	10	1.6	0.1	0.1	0.8	1.1	0.2			
18	S18	04	2.2	1.0	-0.1	2.1	0.0	0.6	2.7		
19	S19	13	0.3	0.8	1.5	0.4	3.3	0.5	0.4	2.7	
20	S20	15	1.6	0.7	0.4	1.1	0.6	0.5	1.8	1.5	0.6

In the table indicating the IRTPRO standardized LD χ^2 χ^2 magnitude for multidimensional IRT, 1 value greater than 10 is found in the correlation between two items. It can be interpreted that local independence is more appropriate for the MDIRT.

Invariance Property

For the assumption of inter-group invariance, the sample is divided into two groups in the unidimensional IRT and the correlation coefficients of the item parameters are calculated from these two groups (Table 6).

In order to examine the invariance property, the correlations of the respondents in the study group for the relationships between the ability parameters divided into two halves are calculated. The fact that most of the correlation coefficients are significant at α = 0.01 level and linear relationships are observed in the scatter plots indicates that the invariance property is met. It is seen in the graphs below that the discrimination parameter has a very low correlation coefficient and a non-linear scatter plot (Figure 1 to 3).

Table 6: Correlation Coefficients for Unidimensional IRT Group 1 and Group 2.

	Groups	*Correlation Coefficient*
Unidimensional IRT	Group1-Group2(b)	0,931**
	Group1-Group2(a)	0,494**

Considering the test information function graph, it can be said that according to the unidimensional IRT model, the groups provides similar information compared to the θ ability levels in the predictions made regarding the ability θ θ levels of the groups throughout the test.

Looking at the graphs above, it is seen that the item characteristic curves do not differ in item discrimination. This means that the probability of answering correctly in items with discrimination differs between low ability levels and high ability levels. In both groups, it is concluded from the graphs that the probability of being successful increases as the ability level increases.

Model-Data Compatibility

Model fit can be evaluated both at the general and item level in IRT. This study has been conducted to determine which

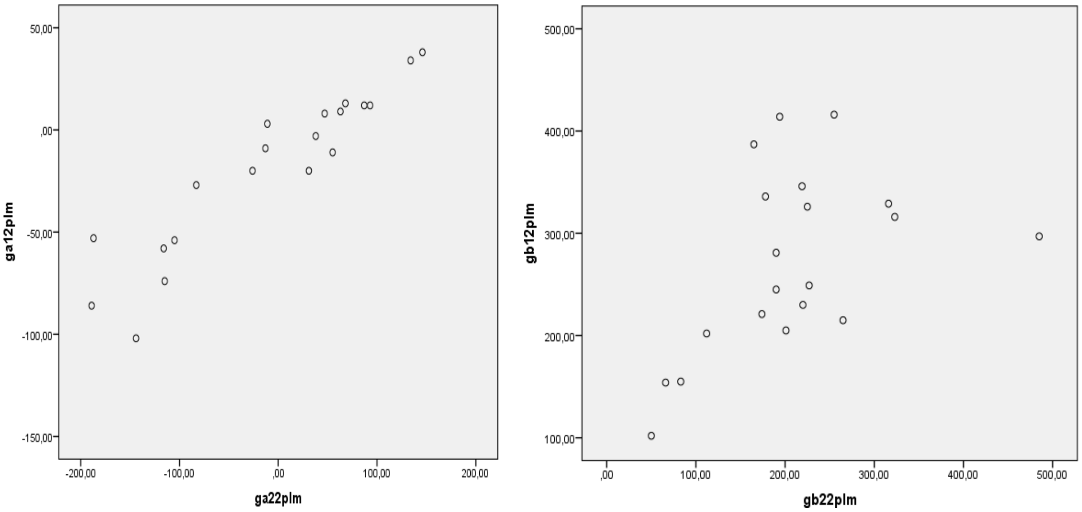

Fig. 1: Scatter Plots of Group1 and Group2 for Unidimensional IRT.

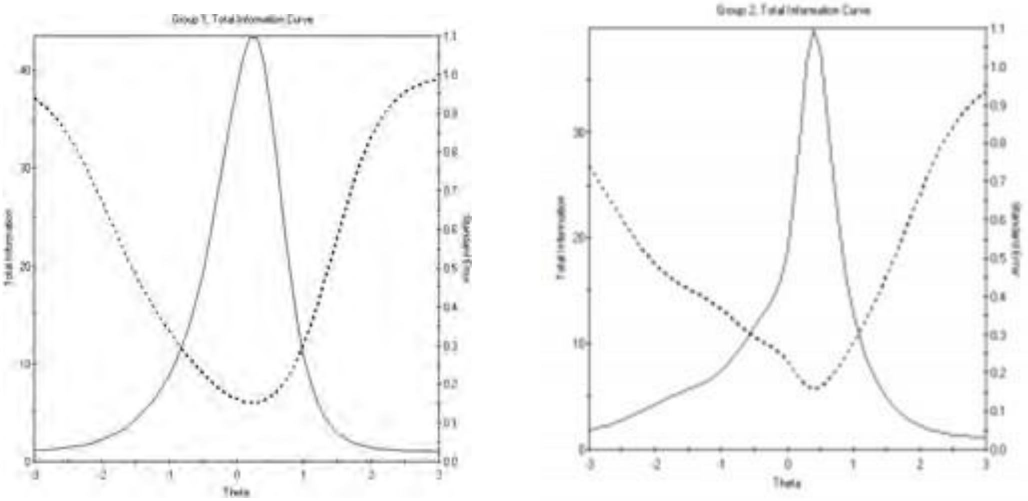

Fig. 2: Test Information Function for Group 1 and Group 2 for Unidimensional IRT.

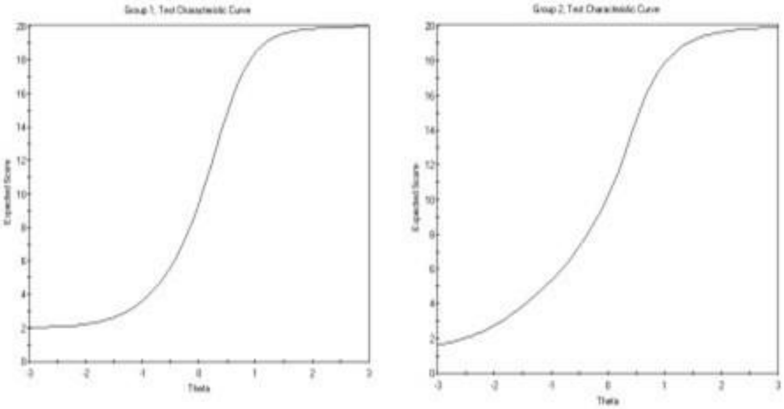

Fig.3: Item Characteristic Curves for Unidimensional IRT Group 1 and Group 2.

model fits the research data better. Therefore, first, the overall model-data fit is evaluated comparatively. There are various methods for assessing the fit of IRT models (Embretson & Reise, 2000). These methods are used to understand how well the model fits the data and provide an important tool for comparing the performance of models along different dimensions. The first and most basic of these methods is the likelihood ratio test - G 2 (LRT) statistic. It takes into account the change in the obtained -2 log likelihood values. The LRT is the difference between the -2 log likelihood values of the models:

$$\Delta G 2 = -2 \ln(L1) - (-2 \ln(L2)) = G1\,2 - G2$$

$L1$ is the maximum likelihood of the first model and the -2ln likelihood value of this model, whereas ($L2$) represents the maximum likelihood of the second model. The degrees of freedom (sd) used to test the significance of ΔG 2 refers to the difference in the number of parameters between the two models being compared. A significant p-value may indicate that the second model (more complex model) with more parameters should be used (Andersen, 1973; Baker and Kim, 2004; Bock and Aitkin, 1981; De Ayala, 2009). In this context, the significance of ΔG 2 determines whether the difference between the models is statistically significant and guides the selection of the more appropriate model.

Other methods used to test model fit include the R^2 statistic comparing various regression models, the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) and the M2 limited information goodness-of-fit statistic. Lower AIC and BIC values indicate a relatively better fit and may indicate that the model fits the data better (De Ayala, 2009). These criteria provide important guidance when comparing between different models and determining which model fits the data better.

Item fit statistics, which are also used in model fit evaluations, are calculated using Bock's (1972) χ 2 index and the results are provided in Table 7. The fit values of the test items of both models are presented and it is checked whether they are significant or not. The fact that the χ 2 value calculated for item fit is significant indicates that the item does not exhibit fitness.

Table 7: Model Comparison.

Statistics based on the loglikelihood	**UDIRT**	**MDIRT**
-2loglikelihood	8646.39	9029.16
AIC	8726.39	9109.16
BIC	8894.97	9279.31

Table 8: Model Comparison.

	Unidimensional IRT	*Multidimensional IRT*
Root Mean Square of Residuals (RMSR)	0.0798	0.0519
Expected mean value of RMSR for an acceptable model	0.0439	0.0498

Overall fit levels are tested to find out which model fit the data better. -2 log, AIC and BIC values calculated for general model fit are compared pairwise. The table indicates the -2 loglikelihood, AIC and BIC values calculated for unidimensional IRT and multidimensional IRT (Table 8).

In model selection, -2 loglikelihood, AIC and BIC values wereconsideredforunidimensional IRTandmultidimensional IRT, but since the number of parameters is equal, $\boldsymbol{G^2 G^2}$ value should be considered. This value is calculated as G2 (1048535) = 3283.23, p = 1 for unidimensional IRT, and since it is not significant, it is thought that the multidimensional IRT model has a better fit (Table 8).

The fact that the Root Mean Square of Residuals (RMSR) value in the upper column of the table above is greater than the Expected mean value of RMSR for an acceptable model indicates that the model of the data set needs to be changed. When looking at the table, it is said that multidimensional IRT is more acceptable because the difference is less (Kelley, 1935).

Item Fit Indices

For unidimensional IRT and multidimensional IRT, the fit values of the relevant test items are presented and the significance level of α = 0.001 is taken as basis. The fact that the χ 2 value calculated for item fit is significant indicates that the item does not exhibit fitness (Table 10).

When the item fit indices are examined in the analysis conducted under the unidimensional IRT model, it is thought that items 4, 13, 18 and 20 might have problems in their fit. In the multidimensional IRT, however, it is thought that items 4 and 1 do not fit well.

To determine whether the models accurately estimated the item parameters, the item parameters and their standard errors are presented below.

In the context of unidimensional IRT, item discrimination (a) and item difficulty (b) standard errors are evaluated in detail. The estimated difficulty parameters under the unidimensional IRT model ranged between b = -1.09 and b = 0.64, and the standard errors for the estimation of these parameters range between (0.05) and (0.17). In contrast to item 13, which is identified as the most difficult item, item 8 stands out as the easiest item. Furthermore, when the discrimination parameter with values between a = 0.83 and

Table 9: Unidimensional IRT and multidimensional IR S-X2 Item Level Diagnostic Statistics.

	Unidimensional IRT			*Multidimensional IRT*		
Item	*X2*	*df*	*Probability*	*X2*	*df*	*Probability*
1	34.57	16	0.0045	27.05	16	0.0408
2	14.48	13	0.3431	14.95	14	0.3835
3	34.37	15	0.0030	28.99	15	0.0161
4	94.60	17	0.0001	100.72	17	0.0001
5	16.00	18	0.5934	21.44	18	0.2569
6	22.39	11	0.0215	22.64	10	0.0121
7	22.81	14	0.0632	24.53	14	0.0394
8	21.56	17	0.2016	23.78	18	0.1618
9	35.24	18	0.0088	39.40	18	0.0025
10	21.16	14	0.0974	23.22	14	0.0566
11	34.58	13	0.0010	22.55	13	0.0473
12	9.61	15	0.8442	18.53	16	0.2929
13	40.98	11	0.0001	44.52	12	0.0001
14	30.84	13	0.0035	31.78	14	0.0043
15	20.77	13	0.0774	15.68	13	0.2661
16	23.45	16	0.1021	26.58	17	0.0643
17	17.44	13	0.1794	17.70	13	0.1689
18	45.86	16	0.0001	38.23	16	0.0014
19	26.61	11	0.0053	23.05	10	0.0105
20	38.87	12	0.0001	30.38	12	0.0024

Table 10: Unidimensional IRT Model Item Parameter Estimates.

Items	*a*	*s.e.*	*c*	*se*	*b*	*se*
M1	2.53	0.25	0.27	0.16	-0.11	0.07
M2	3.61	0.39	0.73	0.22	-0.20	0.06
M3	2.63	0.27	0.10	0.17	-0.04	0.07
M4	1.82	0.19	0.02	0.14	-0.01	0.08
M5	1.18	0.14	0.49	0.12	-0.42	0.10
M6	2.54	0.30	2.66	0.27	-1.04	0.10
M7	2.95	0.32	-0.83	0.21	0.28	0.06
M8	1.33	0.16	1.46	0.15	-1.09	0.13
M9	0.83	0.12	0.75	0.11	-0.90	0.17
M10	3.29	0.37	-1.20	0.24	0.36	0.06
M11	3.40	0.36	-0.40	0.21	0.12	0.06
M12	2.27	0.25	1.81	0.20	-0.80	0.09
M13	3.13	0.40	-2.01	0.30	0.64	0.06
M14	2.62	0.29	1.88	0.22	-0.72	0.08
M15	3.26	0.35	-0.63	0.21	0.19	0.06
M16	2.09	0.27	0.07	0.05	-0.62	0.12

Table 10: Unidimensional IRT Model Item Parameter Estimates.

Items	*a*	*s.e.*	*c*	*se*	*b*	*se*
M17	4.77	0.63	0.05	0.02	0.34	0.05
M18	2.66	0.32	-0.51	0.20	0.19	0.07
M19	4.37	0.55	-2.73	0.39	0.62	0.05
M20	4.72	0.57	-2.02	0.35	0.43	0.05

a = 4.77 is analyzed, it is determined that the items generally have high discrimination. A high a value indicates that the item better discriminates students at low and high ability levels. In this context, items 17 and 20 are established to have the highest discrimination for the test, while the item with the lowest discrimination is item 9 (a = 0.83). When the standard errors of b and a parameters on item basis are examined, it is seen that they have average values ($\bar{X}$b = 0.079 and $\bar{X}$a = 0.325).

In order to examine the invariance property of ability parameters, the correlation coefficient for the relationships between the ability parameters estimated from each of the item sets of the respondents in the study group separated as odd-even is calculated as .863. According to the unidimensional and bi-dimensional IRT models, the invariance property for different item pairs, in which the correlations of the items separated as odd-even are very high, is achieved significantly at the 0.01 level.

In the evaluation of the multidimensional IRT analysis results, standard error averages are used as an indicator of parameter invariance. Data are analyzed without grouping according to dimensions and standard error averages are calculated separately for each dimension to determine how each dimension is affected by parameter invariance (Table 11).

When the standard error means (SHO) of the parameters above are taken, it is seen that the SHO values obtained are between .014 and 2.05 for parameter a1. For parameter a2, this value is between .20 and 1.75. It is possible to say that parameter invariance cannot be maintained for these three conditions. a1 parameter mean is calculated as .664 and a2 parameter mean is calculated as .595. In general, the SHO values for the first dimension ICC a parameter are lower than the SHO values for the second dimension ICC a parameter.

The “c” parameter in this table is the intercept parameter obtained as a result of the interaction of the “b” difficulty

Table 11: Multi-Dimensional IRT Model Item Parameter Estimates.

Item	*Label*	a_1	*s.e.*	a_2	*s.e.*	*c*	*s.e.*
1	S1	4.36	0.67	0.00	-----	0.26	0.77
2	S2	8.67	2.05	0.00	-----	1.45	1.52
3	S3	6.02	1.14	0.00	-----	0.04	1.03
4	S4	0.00	-----	1.94	0.20	0.17	0.37
5	S5	1.35	0.16	0.00	-----	0.51	0.25
6	S6	3.45	0.51	0.00	-----	3.32	0.65
7	S7	0.00	-----	2.71	0.34	-0.89	0.48
8	S8	0.00	-----	1.41	0.21	1.55	0.27
9	S9	0.82	0.14	0.00	-----	0.76	0.17
10	S10	0.00	-----	3.45	0.56	-1.30	0.72
11	S11	0.00	-----	4.21	0.55	-0.49	0.77
12	S12	2.02	0.34	0.00	-----	1.68	0.32
13	S13	0.00	-----	2.95	0.64	-1.95	0.72
14	S14	2.52	0.52	0.00	-----	1.85	0.33
15	S15	0.00	-----	3.08	0.43	-0.55	0.55
16	S16	2.23	0.45	0.00	-----	1.48	0.30
17	S17	0.00	-----	4.18	0.63	-0.92	0.80
18	S18	0.00	-----	2.56	0.34	-0.27	0.45
19	S19	0.00	-----	4.37	1.75	-2.54	1.49
20	S20	0.00	-----	4.62	0.90	-1.57	0.93

parameter and the “a” discrimination parameter and is called the “d” parameter in Multidimensional Item Response Theory. The “d” (intercept) parameter is not related to the feature of having an upper limit, which is rarely encountered in Multidimensional IRT. The “d” parameter mentioned here replaces the difficulty (b) parameter in the One Dimensional ILC and has a different meaning. However, in multidimensional IRT, the d parameter is calculated .

$$d_i = -\sum_{k=1}^{m}(a_{ik}b_{ik})$$

For each item, not as a vector like the a parameter The parameter d in the formula is called the intercept parameter because it consists of the difficulty (position) and discrimination (slope) parameters (de Ayala, 2009).

The table 12 demonstrates the information functions obtained for each item between the -2.8 and 2.8 ability levels. According to this table, the most information is provided at 0.4 ability level with 30.08. It is seen that the standard error of the measurement is inversely proportional to the information function of the test and the item. At the 0.4 ability level, the standard error is also lower than the other standard errors.

The information function in Multidimensional IRT can be considered as an extension of the information function in Unidimensional IRT. This function is represented by adding the direction of information to its mathematical expression. Mathematical expression of multidimensional information;

$$P_i(\theta)[1-P_i(\theta)]\left(\sum_{k=1}^{m}\alpha_{ik}\cos\alpha_{ik}\right)^2$$

In the formula; Pi $\Theta\Theta$............... represents the probability of a respondent at ability level $\Theta\Theta$ to answer item i correctly, while αik .... is the vector representing item i in the latent trait composition (Köse, 2012).. This vector $\Theta_1\Theta_1$ is represented by the angle with the 0x1 axis (Ackerman, 2005)

Table 12: Item Information Functions for Unidimensional IRT Model.

		θ:														
Item	*Label*	*-2.8*	*-2.4*	*-2.0*	*-1.6*	*-1.2*	*-0.8*	*-0.4*	*-0.0*	*0.4*	*0.8*	*1.2*	*1.6*	*2.0*	*2.4*	*2.8*
1	M1	0.01	0.02	0.05	0.14	0.36	0.80	1.39	1.57	1.09	0.53	0.22	0.08	0.03	0.01	0.00
2	M2	0.00	0.00	0.02	0.08	0.34	1.21	2.87	2.86	1.20	0.33	0.08	0.02	0.00	0.00	0.00
3	M3	0.00	0.01	0.04	0.11	0.30	0.73	1.39	1.73	1.26	0.62	0.25	0.09	0.03	0.01	0.00
4	M4	0.02	0.04	0.08	0.17	0.31	0.52	0.74	0.83	0.72	0.50	0.30	0.16	0.08	0.04	0.02
5	M5	0.07	0.11	0.16	0.22	0.28	0.33	0.35	0.33	0.28	0.22	0.16	0.11	0.07	0.05	0.03
6	M6	0.07	0.19	0.48	1.02	1.56	1.47	0.88	0.40	0.16	0.06	0.02	0.01	0.00	0.00	0.00
7	M7	0.00	0.00	0.01	0.03	0.11	0.33	0.90	1.83	2.11	1.27	0.51	0.17	0.05	0.02	0.01
8	M8	0.15	0.23	0.31	0.40	0.44	0.43	0.36	0.27	0.19	0.12	0.08	0.05	0.03	0.02	0.01
9	M9	0.10	0.12	0.14	0.16	0.17	0.17	0.17	0.15	0.13	0.11	0.09	0.07	0.05	0.04	0.03
10	M10	0.00	0.00	0.00	0.02	0.06	0.23	0.75	1.93	2.69	1.68	0.61	0.18	0.05	0.01	0.00
11	M11	0.00	0.00	0.01	0.03	0.13	0.47	1.45	2.78	2.32	0.94	0.28	0.07	0.02	0.00	0.00
12	M12	0.05	0.13	0.30	0.62	1.05	1.28	1.05	0.62	0.30	0.13	0.05	0.02	0.01	0.00	0.00
13	M13	0.00	0.00	0.00	0.01	0.03	0.10	0.35	1.02	2.13	2.31	1.24	0.44	0.14	0.04	0.01
14	M14	0.03	0.08	0.22	0.56	1.18	1.69	1.44	0.79	0.33	0.12	0.04	0.02	0.01	0.00	0.00
15	M15	0.00	0.00	0.01	0.03	0.11	0.39	1.17	2.41	2.38	1.14	0.37	0.11	0.03	0.01	0.00
16	M16	0.04	0.10	0.23	0.49	0.87	1.18	1.11	0.74	0.39	0.18	0.08	0.03	0.01	0.01	0.00
17	M17	0.00	0.00	0.00	0.01	0.06	0.25	1.01	2.82	3.31	1.45	0.39	0.09	0.02	0.00	0.00
18	M18	0.00	0.01	0.03	0.07	0.21	0.53	1.14	1.71	1.51	0.84	0.35	0.13	0.05	0.02	0.01
19	M19	0.00	0.00	0.00	0.00	0.01	0.07	0.30	1.20	3.16	3.19	1.23	0.31	0.07	0.02	0.00
20	M20	0.00	0.00	0.00	0.01	0.04	0.19	0.75	2.32	3.43	1.84	0.54	0.13	0.03	0.01	0.00
Test LInformation:		1.56	2.06	3.11	5.19	8.61	13.3	20.6	29.3	30.1	18.6	7.88	3.29	1.79	1.30	1.14
Expected s.e.:		0.80	0.70	0.57	0.44	0.34	0.27	0.22	0.18	0.18	0.23	0.36	0.55	0.75	0.88	0.94

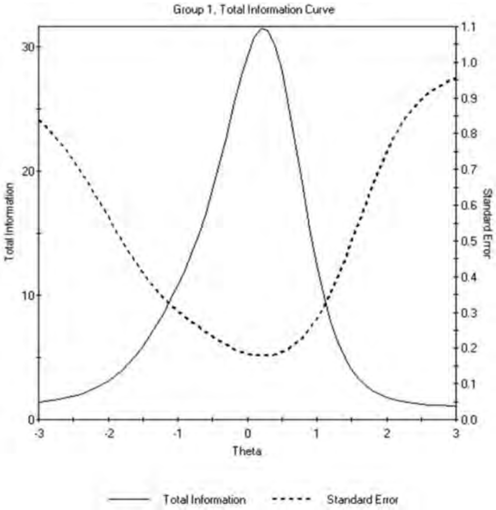

Fig. 4. Unidimensional IRT General Test Information Function

The test information function gives the overall information function of the test according to the ability levels of the test. It is seen that the estimation according to the unidimensional IRT model provides more information between $\theta\theta$=0 ile $\theta\theta$=1. The amount of information provided by the test decreases as one moves to the extremes of the ability scale. The information function is inversely proportional to the standard error of ability estimation. The dashed lines in the graph indicate that the standard error decreases as the amount of information increases.

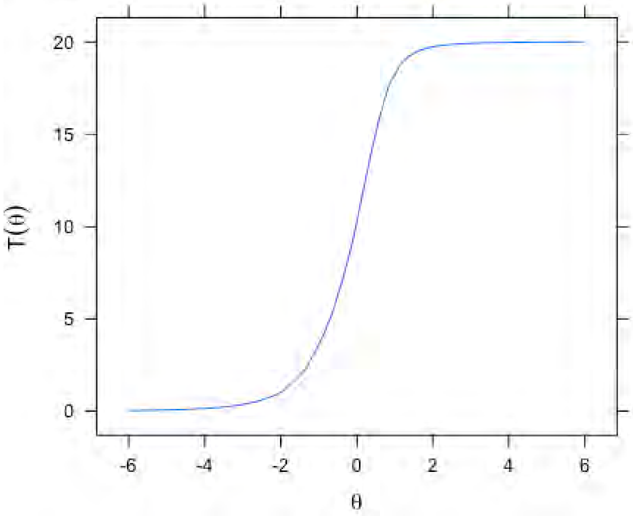

Fig. 5. Characteristic Curve of Unidimensional IRT General Test

Item discrimination is mainly related to the steepness at the midpoint of an item characteristic curve. Steeper curves indicate that the item is more discriminative, while flatter curves indicate that the item has lower discrimination. For items with a flat curve (i.e. items with low discrimination), the probability of answering the item correctly is almost the same at low ability levels and at high ability levels. Respondents' probabilities of success indicate an increasing trend with increasing ability level, which can be seen graphically.

The item characteristic surface is obtained using extended models for multidimensionality. This surface includes the probability of answering correctly and the different θ (theta) planes analyzed on the item, similar to the Item Characteristic Curve (ICC). The item characteristic surface indicates the relationship between an student's level in the relevant ability levels and the probability of answering the item correctly on multidimensional graphical planes. Figure 1.4 indicates the item characteristic surfaces based on the logistic and normal ogive function for an item in a two-dimensional compensated model with α1=0.5; α2=1.5; β = 0 and c=0.2.

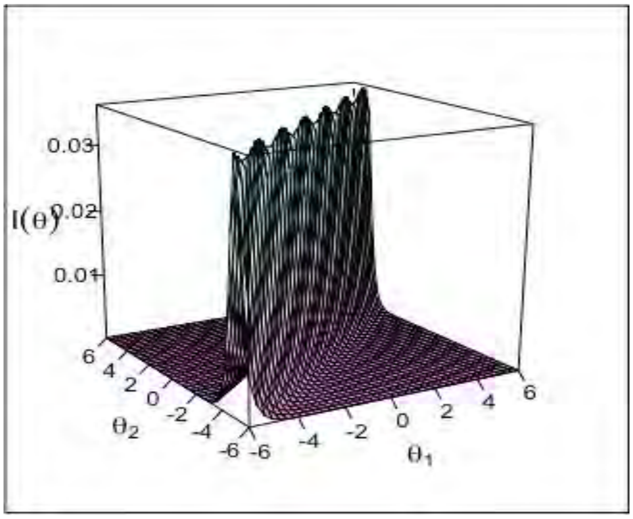

Figure 6: General Test Information Function for Multi-dimensional IRT

The test information function gives the overall information function of the test according to the ability levels of the test. The amount of information provided by the test decreases as one moves to the extremes of the ability scale. The information function is inversely proportional to the standard error of ability estimation. The dashed lines in the graph indicate that the standard error decreases as the amount of information increases.

Examining item characteristic surfaces in the figure, we see that the probability of a correct response increases monotonically with the increase in the elements in the θ vector. Furthermore, it is observed that on these surfaces, a unit increase in θ2 increases the probability of a correct

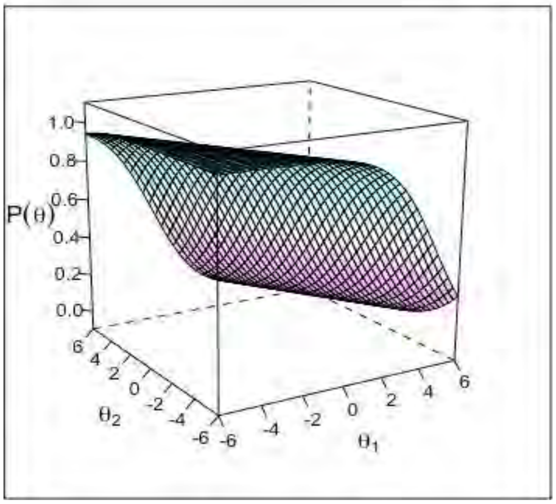

Fig. 7: Characteristic Curve of Multidimensional IRT General Test

response more than the same unit increase in θ1. This suggests that θ2 is more effective than θ1 in increasing the probability of responsing correctly for this particular item.

DISCUSSION

The results of the study indicate that there is no significant difference between the errors of the item parameters obtained fromtheunidimensionalandmultidimensional IRTinthecases of two-dimensional data structures and high interdimensional correlations. However, when the data structure are three and five dimensional, the item parameter errors resulting from the unidimensional IRT increase. Standard errors for item parameters decreased as the sample size increased. When the standard errors arising from ability parameters were analyzed, it was observed that multidimensional IRT estimated with lower errors for all conditions. As a result of the study, it is established that multidimensional IRT provides more accurate results in the analysis of multidimensional constructs, especially in the estimation of student parameters and in making decisions about students.

CONCLUSION

Item Response Theory (IRT) models are frequently used in education and psychology. However, it can be difficult for these models to be completely unidimensional, particularly in cases such as achievement and aptitude tests (Hambleton, Swaminathan, & Rogers, 1991; Reckase, 1997). In cases where the assumption of unidimensionality is not met, applying a unidimensional model to multidimensional test data may lead to errors in ability and item parameters and model misfit. In cases where there is more than one latent trait affecting the respondent, multidimensional IRT models should be used (Ackerman, 1994a-b). Although studies suggest that unidimensional models can be used with moderately multidimensional data, the risk of lack of information and misinterpretation increases in ability estimation with multidimensional data (Drasgow & Parsons, 1983; Hambleton, 1969; as cited in Kreiter, 1993). Discrepancies under unidimensional models are usually caused by the multidimensionality of test data. Research indicates that multidimensional models give more reliable results in the estimation of ability and item parameters when the tests are multidimensional.

According to the findings of this study, student parameter values obtained by multidimensional analyses present lower error rates than unidimensional models. Therefore, it is recommended to prefer multidimensional models when making decisions about students. It is also important for test developers to make predictions by considering the relationships between sub-dimensions. The results of the study demonstrate that multidimensional models estimate student parameters with less error under all conditions.

SUGGESTION

The results of the study suggested that multidimensional models may be more appropriate for assessing students' abilities in comprehensive examinations used at both national and international level.

LIMITATION

Simulation studies have some limitations. These limitations arise from factors such as modelling errors, data scarcity, uncertainties, parameter selection, computer capacity, and validity/generalization. Modelling errors can make it difficult to fully represent real-world phenomena, while data scarcity and uncertainties can affect the accuracy of simulation results. Limitations in parameter selection and computer capacity affect the scope and accuracy of the model, while validity and generalization boundaries determine how applicable the results are in the real world. These limitations should be taken into account when interpreting and using simulation results.

REFERENCES

Ackerman, T.A.(1994a). Graphical Representation of Multidimen-sion-al Item Response Theory Analyses. Paper Presented at the Annual Meeting of American Educational Research Association. New Orleans, LA.

Ackerman, T.A. (1994b). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, 7(4), 255-278.

Ackerman, T. A. (2005). Multidimensional item response theory modeling. In Maydeu-Olivares, A., McArdle, J. J. (Eds.), Contemporary psychometrics (pp. 3–24). Mahwah, NJ: Lawrence Erlbaum.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38(1), 123–140. https://doi.org/10.1007/BF02291180

Baker, F. B., & Kim, S.-H. (2004). Item response theory: Parameter estimation techniques (2nd

ed.). New York, NY: Marcel Dekker.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443–459. https://doi.org/10.1007/BF02293801

Brown, T.A., 2015. Confirmatory Factor Analysis for Applied Research. Guilford publications.

Bulut, O. (2015). Applying item response theory models to entrance examination for graduate studies: practical issues and insights. Journal of Measurement and Evaluation in Education and Psychology, 6(2); 313-330.

Crocker L., & Algina J. (1986). Introduction to classical and modern test theory. New York: CBS Collage Publishing,

Çakıcı Eser, D. & Gelbal, S. (2015). Examining the parameter estimations of simple and complex structured tests with various dimensionality properties based on multidimensional item response theory. Journal of Measurement and Evaluation in Education and Psychology, 6(2); 331-350.

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

Drasgow ve Parsons (1983). Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.

Gül, E & Koç, N. (2017). Examining multidimensional structure in view of unidimensional and multidimensional item response theory. Hacettepe Univesity. Journal of Education, 32(2): 312-326.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H. and Rogers, H. (1991). Fundamentals of Item Response Teory. Newbury Park CA: Sage.

Hasançebi, B., Terzi, Y. and Küçük, Z. (2020). Investigation of Parameter Invariance in Two-Parameter Item Response Theory Model. Journal of Natural and Applied Sciences, 24(2), 438-444.

Hori, K., Fukuhara, H. & Yamada, T. (2022). Item response theory and its applications in educational measurement Part I: Item response theory and its implementation in R. WIREs Computational Statistics, 14(2), 1531.

Kelley, T. L. (1935). Essential Traits of Mental Life, Harvard Studies in Education. vol. 26. Harvard University Press, Cambridge.

Koğar, H. (2014). Comparison of item parameters and model fit obtained from different item response theory applications based on sample size and test length. Unpublished doctoral dissertation, Hacettepe University Graduate School of Educational Sciences.

Köse, İ.A. (2010). Comparison of unidimensional and multidimensional models based on item response theory in terms of test length and sample size. Unpublished doctoral dissertation, Ankara University Graduate School of Educational Sciences.

Köse, İ.A.(2012). Multidimensional item response theory. Journal of Measurement and Evaluation in Education and Psychology, 3(1), 221-229.

Kreiter, C.D. (1993). An emprical iInvestigation of compensatory and noncompensatory test items in simulated and real data. Unpublished Doctoral Dissertation. The University of Iowa.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading MA: Addison-Welsley Publishing Company.

McDonald, R.P. (2000). A basis for multidimensional item response theory. Applied Psychological Measurement, 24, 99.

Muthén, B., & Kaplan D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. British Journal of Mathematical and Statistical Psychology, 38, 171-189.

Reckase, M.D. (1997). Models for multidimensional tests and hierarchically structured training materials. Technical Report. The American Colage Tesing Program. Iowa City, Iowa.

Reckase, M. D. (2009). Multidimensional item response theory (Statistics for social and behavioral sciences). New York: Springer.

Sünbül, Ö. Ve Erkuş, A. (2013). Examining item parameter invariance for several dimensionality types by using unidimensional item response theory. Mersin University Journal of the Faculty of Education, 9(2), 378-398.

Yakar, L. (2017). Retrofitting of cognitive diagnosis and multidimensional item response theory models. Unpublished doctoral dissertation. Hacettepe University Graduate School of Educational Sciences.

Yavuz, G. (2014). Comparative analyses of multidimensional item response theory models and software. Unpublished doctoral dissertation. Hacettepe University Graduate School of Educational Sciences.

Yılmaz Koğar, E, & Çakıcı Eser, D. (2015). Tek ve Çok Boyutlu Madde Tepki Kuramına Dayalı Bir Veri Analizi Yazılımı: IRTPRO 2.1. Journal of Measurement and Evaluation in Education and Psychology. https://doi.org/10.21031/epod.29597