Examining the Two Categorical Datas by Jmetrik, Bilog-Mg and Irtpro with Application of Mathematics Exam

Gokhan Aksu, (Instructor)
Adnan Menderes University/Turkey
Cigdem Reyhanlıoglu (PhD Candidate)
Hacettepe University/Turkey
Mehmet Taha Eser, (PhD Candidate)
Hacettepe University/Turkey

Doi: 10.19044/esj.2017.v13n33p20 URL:http://dx.doi.org/10.19044/esj.2017.v13n33p20

Abstract

The aim of this study was to examination of two-category rated mathematics course final exam based on Item Response Theory data analyzed with the help of 2-Parameter Logistic Model and determination of the ability and standard errors with the help of different programs. This study involves a comparative interpretation of some descriptive statistics and analysis. Therefore, research has characterized as relational model which is one of the general survey models. For this purpose, 771 students' final achievement test responses to a 20-point final exam, were analyzed by BILOG, IRT PRO and JMETRİK programs. Item Response Theory assumptions were analyzed with SPSS and Factor 9.3 programs. Working as a result of the analysis of data all of the IRT assumptions are met and the most appropriate model of data set has been concluded that the two-parameter logistic model. The study also found that there is a statistically significant relationship between the estimated parameters related to individual ability and error at the level of .01. Especially compared to the others there is also significant relationship between JMETRİK and IRT PRO. Different models and methods of research proposals have been made in terms of response patterns to be analyzed a gain for the same data set.

Keywords: BILOG, IRTPRO, JMETRIK, Ability Parameters, Error Parameters

Introduction

From past to present, there have been many studies focusing on "What and how should we assess?" in the assessment and evaluation phase

of the education, and in the wake of the results of these studies, important theories were put forward. The first theory was Classical Test Theory (CTT), used the individuals' grades gained from exams and made some calculations. Second theory Item Response Theory (IRT) aims to predict the ability levels of the individuals using statistical methods. Although it has many theoretically weak sides, CTT has wider study area than IRT (Hambleton and Swaminathan, 1991). The most important reasons for this are that CTT has less assumptions, and that its assumptions are met easily and its parameter are predicted more easily than IRT. However, as of 21th century, the popularity of IRT has risen. We cannot deny IRT's the role of estimating on individual base not group base in the rise of this popularity. As the purpose of the study is one of the most important factors that help the researcher to determine which theory to use in his study; appealing assumptions of the IRT has a big role in raising its importance rapidly.

Two categorized (rated in 0-1) IRT models that are seen the most frequently in the literature are; logistic model (Rash Model), 2-parameter logistic model and 3 -parameter logistic model (Hambleton and et. al., 1991). 2 categorized models are rated as (1,0). The correct answer is coded1, and the wrong answer is coded 0. For example, two categorized models can be applied to multiple choice questions and true-false tests. It is impossible to rate as (1,0) in multi-categorized models. There is not one correct answer for the items that constitute the assessment tools that multi-categorized models can be applied. For example, multi-categorized models can be applied in the written exams (Zheng and Rabe-Hesketh, 2007). The primary multi-categorized IRT models are; Graded Response Model (GRM), Modified Graded Response Model (M- GRM), Partial Graded Model (PCM) and Generalized Partial Credit Model (DPCM) (Embretson and Reise, 2000).

The assessment tool used in the scope of our study is rated as 2 categorized. Therefore brief information about 2 categorized logistic model types was given in the rest of the study.

The logistic model type in which all the items in an assessment tool has equal discriminating power (a), chance parameter is assumed to be low and the same for all items, and item difficulty parameter takes different values is 1PLM, i.e. Rasch Model. The logistic model in which items' item discriminating powers (a) and item difficulty parameters (b) in an assessment tool can be different, but the chance parameter (c) is assumed to be low and the same for all items is 2PLM. The logistic model type in which discriminating power, difficulty parameters and chance parameters of the items are different is 3PLM. All these three model types have advantages and disadvantages (Hambleton and Swaminathan, 1985).

In the cope of the study; related data set was analyzed through IRT assumptions and in the framework of model-data concordance, and

appropriate logistic model was determined, and findings gained after this logistic model was applied to the data set were interpreted. In the last phase of the study; the logistic model was applied to the data set through three different computer programs (BILOG-MG, IRTPRO, JMETRIK), and correlations between the predictions gained from the programs for the ability levels of the individuals were analyzed, and through this it was aimed to figure out which programs were giving algorithmically similar results. Also, it was tried to find out which computer program is likely to be used in the future studies by the researchers. The second phase of the study is also a validity study for the first phase of the study. Briefly, the real purpose of the study to inform the readers if BILOG-MG, IRTPRO, JMETRIK programs, which are thought to be popular, and observed to be used a lot in the studies about IRT, give similar results in terms of ability parameter predictions.

When the related literature is analyzed, even though there are many studies analyzing model-data concordance in two categorized data sets, there is no study in which ability level of the individuals in two categorized data sets are compared through different computer programs. As a result of this study, it is though that the similarity or the difference in the results of these three computer programs will help the researchers in choosing the program in the future when they apply IRT models to be used two categorized data set through one computer program.

Güler, Uyanık and Teker (2014); defined a group of 1250 people via random sampling from 5989 people to whom a multiple choice Turkish test was applied, and data set of the related group was used. According to the study results, the highest correlation in terms of item difficult indices (0,99) was between CTT and 1PLM, and the highest correlation in terms of item discrimination parameters was between CTT and 2PLM. Although, 3PLM was seen as the most appropriate model for data-model concordance, the model rendered the lowest correlation with CTT was 3PLM.

Huang and Others (2013) analyzed an exam consisting of 50 questions and was applied to 170 students and rated through two categorizations. They analyzed the data set of that exam through 1PLM and 2PLM. They concluded that 1PLM was more appropriate for the data set as there was no discordant item, the reliability of the test was 0,81 and fit indices (AIC and BIC) are favor for 1PLM.

Nenty and Adedoin (2013); defined a group of 10,000 people via random sampling from 36,939 people to whom a multiple choice mathematics test was applied, and data set of the related group was used. The researchers calculated the item parameters of the related data set through CTT and IRT (2PLM-3PLM). They analyzed the significance item parameters for dependent samples via t test- in this calculation invariance concept was also considered- and it was observed that there was not a

statistically important difference in the item parameters in the framework of 2 different theories. Also, it was observed that there was not a significant difference in terms of item difficulty parameters between 2PLM and 3PLM. Uyanık, Kaya and Güler (2013) in order to determine the best model for the data set, they took the number of items that were appropriate for the model and chi-square test results as a basis. They concluded that among 1 PLM, 2PLM and 3PLM, 2PLM was the best model for PISA 2009 mathematics subtest.

Weiss and VonMinden (2012) applied among 2PLM and 3PLM through Xcalibre 4.1 and Bilog-MG programs and compared the results with each other. Correlations between item parameters and approximate values of average square root error were analyzed, and according to Xcalibre 4.1 is a more appropriate program for the data set than Bilog-MG.

In this study, it was aimed to analyze mathematics exam data though 2 categorized logistic models of IRT, to determine the appropriate logistic model for data set and to interpret the statistical results related to the determined logistic model. In the second phase of the study, 2 categorized logistic model correlations of ability level predictions and standard error parameters of ability predictions were gained from 3 different computer programs related to 2 categorized logistic model types, which are compatible to the data set. In the result part, it was aimed to compare and interpret these correlations and standard error parameters. As the real data set was used and considering which one of the three computer programs was more appropriate for 2 PLM in terms of 2 categorized data set, the researchers were provided a comparative situation in terms of preference.

Methodology

The study contains the analysis of related data set through certain 2 categorized logistic model methods and computer programs and comparative interpretation of certain descriptive statistics related to the data set and results of the analysis. For this reason, it has relational screening model from general screening models.

The sampling of this study is composed of 771 students studying at Aydın Adnan Menderes University, Main Campus. The students were asked the final exam of the mathematics, which is a compulsory course. Basic Mathematics Course final exam questions were analyzed. Two different academicians working at the same institution and three different mathematics teachers working at different high schools provided help for the sake of expert view. Necessary arrangements were done after the expert views were taken, and the final draft of the 20-question exam was prepared. While preparing the final exam questions, the questions asked before by OSYM (Student Selection and Placement Centre) in the DGS, KPSS and ALES

exams were analyzed in terms of the topics. And these questions' numbers were changed and used in the exam. In preparing the exam, questions, which were defined as medium difficulty by the experts, were tried to be used. According to the information taken from Student Affairs Centre, there were 1218 students registered to the Basic Mathematics Course. As the exam was not a speed test, and to allow all the students see all the items in the exam, student were given 40 minutes for 20 questions. Considering that OSYM gives 1 minute for each question in the exams, the exam duration of this study was thought to be appropriate. Students marked their answers to the optical reader for the questions with five multiple choices. The answer sheets were converted into .txt format so that the data could be analyzed. The correct answers were coded as 1 and the wrong answers were coded as 0. To analyze the data BILOG, SPSS, Factor 9.3 and for local independency IRTPRO programs were used. In the second phase of the study, different package programs (BILOG-MG, IRTPRO, JMETRIK) were used to analyze the error parameters and ability levels of the individuals. The reason to choose these package programs was that BILOG-MG is a program that can make analysis for IRT related to only two categorized data. IRTPRO is a paid program that can make unidimensional or multi-dimensional IRT analyses. And also, JMETRIK is a open source and free program. The difference of JMETRIK from other open source programs making analyses for IRT is that it has link on its interface for IRT analysis related to two categorized data sets and analysis for CTT.

The mean, confidence interval, variance, skewness and kurtosis values of 20 items are shown in Table 1.

Table 1. Descriptive statistics and item statistics for test items

Item	Mean	Variance	Skewness	Kurtosis
1	0.419	0.243	0.329	-1.891
2	0.313	0.215	0.810	-1.344
3	0.331	0.221	0.720	-1.480
4	0.284	0.203	0.959	-1.080
5	0.383	0.236	0.484	-1.765
6	0.402	0.240	0.400	-1.839
7	0.333	0.222	0.708	-1.498
8	0.270	0.197	1.039	-0.921
9.	0.297	0.209	0.890	-1.208
10.	0.431	0.245	0.281	-1.920
11	0.431	0.245	0.281	-1.920
12	0.328	0.220	0.733	-1.462
13	0.300	0.210	0.876	-1.232

14	0.300	0.210	0.876	-1.232
15	0.409	0.242	0.373	-1.860
16	0.416	0.243	0.340	-1.883
17	0.329	0.221	0.727	-1.471
18	0.309	0.213	0.829	-1.312
19	0.304	0.211	0.856	-1.267
20	0.416	0.243	0.340	-1.883

There are descriptive statistics item statistics in table 1. Descriptive statistics item statistics are important in terms of having information about grade distribution and quality of the items. As seen in table 1, the percentage of giving correct answer to 20 items is between 0,270 and 0,431. This means that most of the samples gave incorrect answer to the question 8. Approximately half of the individuals answered correctly to the questions 10 and 11. When item difficulty coefficients and item variance are analyzed, it is seen that these values are between 0,197 and 0,245 (Turgut and Baykul, 2012; 226). When the maximum value of the item variance is thought to be 0,250 (p=0,50), it can be said that the items in the test reveals the differences of the individuals in terms of the assessed feature (Baykul, 2010; 262). When skewness and kurtosis coefficients were analyzed, it can be said that the items between the ±1 range, have normal distribution (Atılgan, Kan and Doğan, 2013). When skewness coefficients are analyzed, it is seen that only the question 8 has a little higher value than the desired value. This means that, the grades of the students are between wide ranges. As the students in the sampling have different grades, we can interpret that 'the group is heterogenic and the frequency of the grades is low".

Data Analysis

Analysis of Item Response Theory Assumptions

First of all, in the Item Response Theory (IRT) some assumptions must be met before the defined model is used. According to Köse (2010) quoted from Spencer, if these assumptions are not met, there will be problems in interpreting the results and choosing the model. Unidimensioned IRT has three widely accepted assumptions. These are; unidimensionality, local independence, and being a speed test or not (Hambleton and at al., 1991; Hambleton and Swaminathan, 1985). After the assumptions are met, invariance principle must be analyzed. According to IRT, there must be relation between the individuals' skills and features that are not directly observable in a specific field or the answers to the question items that test this field, and this relation can be defined mathematically (Rupp and Zumbo, 2006; Hambleton and et. al, 1991; Mckinley, 1989).

One of the most important assumptions of IRT is that all items assess the same ability or the same ability sets. However, in many assessments, the test items individually can assess different ability or ability sets. For this reason, it is necessary to evaluate if the test is unidimensional or multidimensional. Stout (1987), developed a linear factor analysis method for nonparametric hypothesis in order to identify the dimensionality of a test data set. However, analysis can be done through data that met necessary assumptions for factor analysis. One of the two important assumptions of factor analysis is normality and the other is size of the sample. If these two assumptions are met can be determined by Kaiser-Meyer-Olkin (KMO) coefficient and Bartlett sphericity test. KMO test is on one of the criteria to test data structure for factor analysis in terms of sample. KMO is a test that compares the size of observed correlation coefficient and partial correlation coefficient (Kalaycı, 2010). That KMO value is high means that all variables in the scale are predicted perfectly by the other variables. (Çokluk and et.al., 2010). According to Çokluk and et.al. (2010) quoting from Leech, Barrett and Morgan (2005) if the scales used for KMO in terms of the size of sampling are between 0,00-0,50 analysis cannot be done; 0,50-0,60 is "bad", between 0,60-0,70 is " poor", between 0,70-0,80 is "fair", between 0,80-0,90 is "good", and over 0,90 is excellent.

Multivariable normality is the state that all variables and all linear combinations of all the variables are distributed normally (Tabachknick and Fidell, 2001). That the data is multivariable because normal distribution is determined by Barlett's Test of Sphericity. The higher Barlett's Test of Sphericity result is, the higher the possibility of result to be significant (Tavşancıl, 2005). Barlett's Test of Sphericity renders chi square test. As in all chi square tests, significance value is looked for in this test. If the value is lower than the significance level, it is understood that the result is different from r correlation or unit matrix in covariance matrix. This means a factor can be emitted from the correlation matrix. KMO and Bartlett test results gained from SPSS.15 statistics program is given in Table 2.

Table.2 KMO and Bartlett's test

KMO		0,792
Bartlett's Test of Sphericity	X	1397,787
	df	190
	Sig.	0,000

As seen in Table 2, the statistic gained for KMO sampling efficiency is 0,792 which is accepted a "good" to conduct the analysis. Chi Square test for Barlett's Test of Sphericity is significant. These results mean that data shows a normal distribution. A factor analysis was conducted in order to determine if the data has single or multi factors. To determine if a data set

has single or multi factors, related proofs must be revealed clearly. According to Lord (1980) that the items have high load value in the first factor, and that while eigenvalue of the first factor and the variation that it explains is high, the same value in the second factor is low, and that there is a proximity between the eigenvalues of the second factor and the consecutive factor point unidimensionality. The eigenvalue gained for the data set used and explained variance amounts are shown in Table 3.

Table 3. Total variance explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,302	16,509	16,509	3,302	16,509	16,509
2	1,444	7,221	23,730	1,444	7,221	23,730
3	1,204	6,018	29,748	1,204	6,018	29,748
4	1,104	5,522	35,270	1,104	5,522	35,270
5	1,044	5,219	40,490	1,044	5,219	40,490
6	1,012	5,061	45,551	1,012	5,061	45,551
7	0,999	4,995	50,546			
8	0,960	4,799	55,345			
9	0,927	4,635	59,981			
10	0,884	4,421	64,401			
11	0,851	4,255	68,657			
12	0,817	4,087	72,743			
13	0,808	4,042	76,786			
14	0,766	3,831	80,617			
15	0,735	3,674	84,290			
16	0,697	3,483	87,773			
17	0,670	3,352	91,125			
18	0,634	3,170	94,295			
19	0,608	3,039	97,335			
20	0,533	2,665	100,000			

As seen in Table 3, the numbers of the items are as many as the numbers of components. The first column under the Initial Eigenvalues title, total eigenvalue (Total) in terms of each factor's contribution to total variance, the percentage in terms of contribution to total variance (variance %) and Cumulative percentage in terms of contribution to variance (Cumulative %) are given. And under the title of Extraction Sums of Squared Loadings, the numbers of factors and suggestion to which component can be accepted as factor are given. As seen under this title, five factors are suggested. The reason for suggesting five factors is that there are 6 components with eigenvalues over 1. It is seen that 6 factors' contribution to variance is 45,55%. However, the important point while determining the

number of the factors is the importance of the contribution of the each factor to the variance (Çokluk and et al 2011). When the (Variance %) values are analyzed under the title of Initial Eigenvalues, it is seen that the first component has a great contribution to the variance while the other five components have low contributions. In such a case, it can be decided to define the factor number as 1, however to make this decision scree plot must be analyzed. The scree plot constructed after the analysis is seen in Figure 1

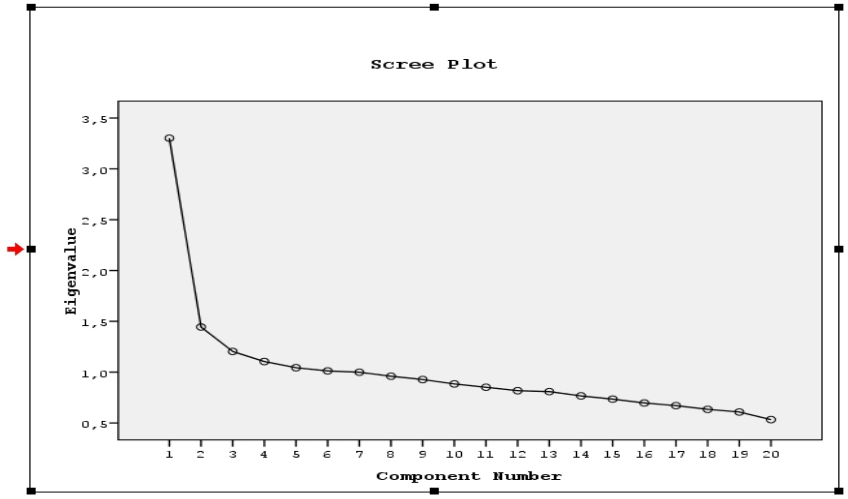

Figure 1. Scree plot

As seen Figure 1, the components in the axis y, go down towards the axis x. This downslope is shown with the dots in terms of their contribution to the variance. Each space between two dots means a factor. As seen in the figure, after the second dot the slope forms a plateau. The components' contributions to the variance after the second dot are low and approximately the same. For this reason, the number of the factor is thought to be 1.

Although SPSS.15 package program renders a unidemensional data structure, indeed it doesn't make a factor analysis that is based on tetrachoric correlation analysis in conformity with the aim of the study (Çakıcı Eser, 2013). SPSS.15 package program make a factor analysis based on Pearson correlation matrix. For this reason, to reveal the dimension of the two categorized structure, parallel analysis made by taking tetrachoric correlation analysis as a base technique was conducted using FACTOR 9.3 and the results gained are seen in Table 4.

Table 4. Results of parallel factor analysis

Variable	Real-data % of variance
1	28.1
2	9.6
3	7.7

4	6.5
5	6.0
6	5.9
7	5.2
8	4.4
9	4.2
10	4.1
11	3.6
12	2.9
13	2.6
14	2.4
15	2.3
16	1.9
17	1.3
18	1.1
19	0.5
20	0.0

**Advised number of dimensions when 95 percentile is considered: 1

When the variances of variables gained through the parallel analysis based on FACTOR 9.3 program based tetrachoric correlation analysis are analyzed, it is seen that the variance rate related the first variable is 28%, this rate after the second variable goes down beginning with 9% after the second variable. In this context, it can be said that the structure is unidimensional. Moreover, the dimension number suggested under the table is defined as 1 in the 95% confidence interval. This shows a consistency with the number of dimension decided according to the variance rates. Besides, this result shows a consistence with the factor analysis that is based on pearson correlation analysis and SPSS.15 package program. As a result, according to the two different factor analyses results, it can be said that unidimensionality which is one of the important assumptions of IRT is met.

The second assumption of IRT, the local independence is that the responses of the individuals to the items are statistically independent or unrelated when the ability, which affects the test performance, is kept the same (Lord and Novick, 1968; Hambleton and the others, 1991). Although to meet the unidimensionality assumption, the items in the test must be related; in this assumption the items must be independent for a specific level of ability. In the local independence assumption, relation between the items and independency are analyzed under a specific ability condition. Moreover, meeting the unidimensionality assumption is generally seen enough to met the local independency assumption. However, in this part of the study the local independency assumption is tested, as well.

In recent years, various indexes emerging from situational covariance are developed in order to assess if the local independency assumption is met

or not. In this study, to test the local independency, the "student" version of IRTPRO statistics package program was used. To make analysis in this version of the program, some conditions must be met. The maximum item number is 25 and maximum individual number is 1000 in the date file to be analyzed. Also, the data structure can be maximum 3 dimensional. As the data set in this study meets these conditions, the local independency assumption was tested through IRTPRO statistics package program. The results concerning whether the items meet the local independency assumption is seen in Table 5.

Table 5. Marginal fit (X^2) and Standardized LD X^2 statistics for group 1

Item	X^2	1	2	3	4	5	6	7	8	9	10
1	0.1										
2	0.1	2.4									
3	0.1	0.2	1.2								
4	0.1	-0.3	-0.5	-0.3							
5	0.2	-0.6	-0.6	-0.1	2.6						
6	0.1	-0.5	-0.7	6.6	0.2	2.0					
7	0.1	0.3	-0.7	-0.7	-0.6	-0.5	2.6				
8	0.1	0.7	4.7	-0.5	-0.4	-0.5	7.6	11.1			
9	0.1	0.2	0.1	-0.3	0.3	-0.4	-0.3	-0.7	-0.1		
10	0.1	-0.6	-0.5	-0.6	2.3	-0.4	-0.1	-0.6	1.6	3.0	
11	0.1	-0.6	1.2	-0.5	-0.5	2.0	3.8	0.1	1.0	-0.5	-0.1
12	0.1	-0.6	-0.4	2.3	-0.1	-0.6	-0.7	-0.2	-0.7	1.4	0.6
13	0.1	3.5	0.0	-0.5	-0.6	-0.6	2.2	0.3	-0.7	-0.6	1.3
14	0.1	-0.3	-0.4	-0.7	-0.1	6.5	0.9	-0.1	5.1	-0.5	0.1
15	0.2	-0.5	-0.6	-0.5	-0.4	0.1	-0.0	-0.1	0.0	0.2	2.3
16	0.1	4.7	-0.5	0.5	-0.5	-0.6	1.2	-0.7	-0.7	-0.6	-0.3
17	0.1	-0.5	-0.4	-0.7	0.1	-0.2	-0.6	-0.7	-0.2	-0.4	-0.0
18	0.1	2.4	-0.7	0.9	3.5	4.7	0.2	0.1	1.6	-0.7	-0.6
19	0.1	-0.6	-0.5	-0.2	-0.5	-0.2	-0.4	-0.4	-0.6	0.3	-0.2
20	0.1	-0.1	-0.7	-0.4	0.0	0.1	-0.1	-0.3	-0.6	-0.6	-0.6

Item	X^2	11	12	13	14	15	16	17	18	19
11	0.1									
12	0.0	4.1								
13	0.0	-0.4	4.3							
14	0.0	-0.4	1.5	-0.3						
15	0.2	-0.3	-0.3	1.5	3.2					
16	0.0	0.7	-0.7	6.4	-0.7	0.1				
17	0.0	-0.4	-0.6	0.9	0.1	1.0	6.3			
18	0.0	0.6	3.0	1.6	-0.0	-0.6	8.6	2.7		
19	0.1	1.1	-0.5	-0.3	-0.6	1.4	2.2	-0.6	5.8	
20	0.1	-0.6	0.1	-0.3	-0.6	-0.6	-0.5	-0.3	-0.5	3.5

According to IRTPRO package program handbook, LD χ^2 values over 10 point out local independency. As seen in Table 5, the number of item pair whose LD χ^2 value ise over 10 i.e. local independent is only 1. The local independency assumption is thought to be met because LD value is not over 10 and only one item accesses the desired value, and also the unidimensionality is not met.

The third assumption of IRT, Analyzing the Situation of being a Speed Test or not. According to the analysis of the forms processed through optical reader, no student of 771 left an item blank because 2 minutes is given to for each question in success test with 20 items prepared in the scope of final exam. Also, bearing in mind that the students give importance to the final exam, which affects their passing with 70%, this result is expected. As Hambleton and Swaminathan (1985) point out, the rate of the individuals who completed the test gives information whether the test is a speed test or not. When the answer patterns for final exam of all individuals are analyzed, it is seen that all 771 individuals marked the last item of the test. For this reason, it is concluded that the success test with 20 items is not a speed test.

Analysis of Model-Data Concordance

In the analysis of data with two answers (1,0) like success tests, BILOG MG program is used. In this study, to analyze model-data concordance respectively 1LP, 2LP, and 3LP models were used and -2 LogLikelihood (-2LL) values were gained for each model. -2LL values were gained for each model, are shown in Table 6.

Table 6. -2 Loglikelihood values for inter models

1PLM	2PLM	3PLM
-2 Log Likelihood: 18912.892	-2 Log Likelihood: 28829.682	-2 Log Likelihood: 18860.011

As seen in Table 6, to determine which model is appropriate for the data structure, in the degree of freedom, the difference between -2LL values is analyzed if it is over than the desired value by looking at the χ^2 table. As there are 20 items in the test, p=0,01 and sd=20 and desired value for χ^2 is 37,57. For 1PL and 2PL models, the difference between -2LL values is 83,21. As the value gained is over than the intended value, it is determined that 2PL model is more appropriate for data structure than 1PL model. Also, the difference of -2LL values of 2PL and 3PL models is 30,33. As 20 slope (a) parameters are added to each item in 2PL model, the freedom degree was calculated as 40. Inχ^2 table, for p=0,01 and sd=40 the intended value is 63,69 and for 2PL and 3PL models the difference between -2LL values is

30,33, and value gained doesn't exceed the intended value. For these reasons, 2PL model is thought to be more appropriate for the data. However, as stated in IRTPRO handbook, the model with the lowest -2LL value among the models is the most appropriate for model-data concordance, and as the 2PL model whose -2LL is the lowest among the models given in Table 6, it was decided to use 2PL model to analyze the data.

Examining The Ability Parameter Invariance

In order to analyze ability parameters' invariance, the items in the test were dived into two groups which are formed randomly as form X (1-3-6-8-9-12-13-15-18-20) and form Y (2-4-5-7-10-11-14-16-17-19), and the correlation between these forms were calculated. As calculating the correlation just in two groups was not enough, the items in the test were grouped again as even and odd numbers. By doing this 4 different rating types are gained as form x, form y, odd numbered items and even numbered items. For correlation analysis, data file related to forms for BIOG MG program is shown in Figure 2, and the titles of these data files were defined as; **Form X, Form Y, Odd Questions Even Questions.**

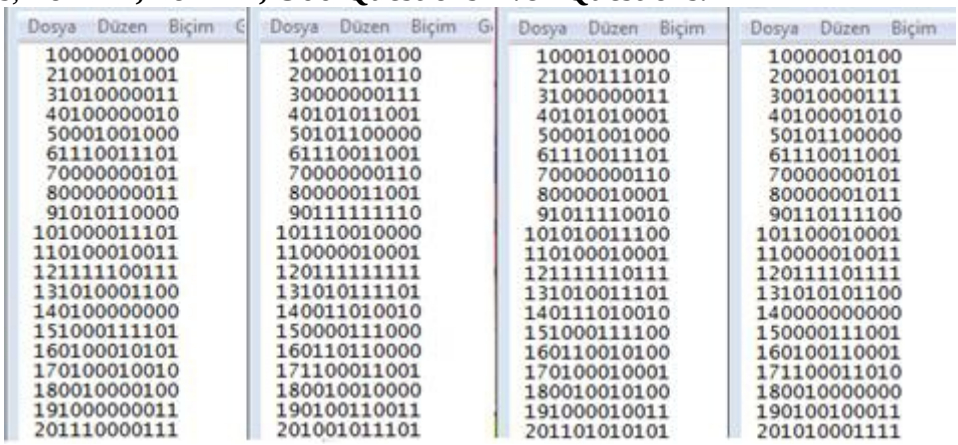

Figure 2. Data files for form x, form y, odd questions and even questions

The first three columns in the data file gives the ID numbers of the individuals, the columns from the 4th column to 13th column the answers of the individuals to the items are shown as being wright or wrong and patterned as 0-1. There is no space left between ID and answer pattern and data was formatted as (3A1, 10A1). Ability parameters gained in the result of analysis made according to 2PL model in BILOG program, were tested through correlation analysis. The ability parameters of the individuals were made appropriate for the analysis and then analyzed in SPSS program, and the results are shown in Table 7.

Table 7. Results of pearson correlation

		form_x	form_y	Even num.	Odd num.
form_x	Pearson Correlation	1	,566(**)	,698(**)	,815(**)
	p		,000	,000	,000
	N	771	771	771	771
form_y	Pearson Correlation	,566(**)	1	,797(**)	,769(**)
	p	,000		,000	,000
	N	771	771	771	771
Even num.	Pearson Correlation	,698(**)	,797(**)	1	,539(**)
	p	,000	,000		,000
	N	771	771	771	771
Odd num.	Pearson Correlation	,815(**)	,769(**)	,539(**)	1
	p	,000	,000	,000	
	N	771	771	771	771

As seen Table 7, the values gained from correlation analysis were found to be statistically significant. When the size of the correlation coefficients are analyzed, it is seen that there are positive medium and high-level relations between different forms. According to this result, it can be said that ability parameters predicted for students are similar to each other with the help of different forms. Accordingly, it can be said that for 2PL model ability parameters meet the invariance assumption.

In order to prevent the students sitting next to each other or one after the other, the final exam containing 20 questions were mingled and divided into A, B, C, D forms. Although the number of the individuals to analyze for each form was 189-196, data files related to each form analyzed in order to see the relation among the parameters predicted for individuals for these forms. For correlation analysis, raw data file was saved with .prn extension in different names for BILOG MG program after the answers for A, B, C, D booklets were filtered and divided in Excel program; and relations among the ability parameters predicted related to the individuals for 4 different booklets were analyzed.

Table 8. Correlations between predicted ability parameters for group A-B-C-D

	Group A	Group B	Group C	Group D
Group A				
Group B	0,123			
Group C	0,146*	0,155*		
Group D	0,064	0,076	0,045	

The numbers shown with * in table 8 point out the significance of the correlation values. As seen in the table, only two of the correlation values are significant, but the others are low level. Although the invariance assertion of

the ability parameters of the individuals was verified when the answers of 771 individuals were formed as odd-even and formx-formy; when we group students as A, B, C, D; this assertion isn't verified. One of the fundamental reasons for this result is thought that the number of the individual for each group is averagely 190. However, when bearing in mind that the number of the individuals must be 1000 and over for IRT, the violation of this assumption is though to be an expected result.

Examining The Item Parameter Invariance

To analyse this assumption, the individuals were divided into two groups in terms of their ID numbers' being odd and even. Also, the individuals were divided into 27% lower groups and upper groups in terms of their predicted abilities through 2PL, and correlations between a and b parameters for each group. 208 students in the lower groups and upper groups and the data files formed in order to predict parameters for the groups created considering the ID numbers being odd or even are shown in Figure 3.

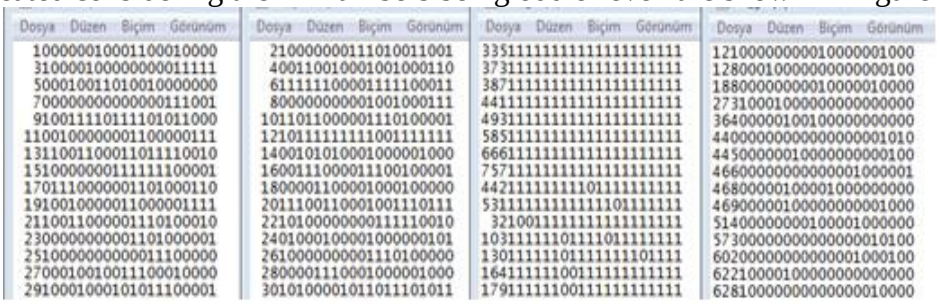

Figure 3. Data file for item parameter invariance

Data files whose one part is seen in Figure 3, were analyzed through BILOG MG and, a and b parameters were calculated for each item. The results of the analysis made to find out the direction and severity of the relation between the parameters gained are shown in Table 9.

Table 9. Correlations between item parameters

2PL Model		
	a parameter	b parameter
Odd ID and even ID students	0,643**	0,907**
Lower group and super group students	0,683**	0,712**

When the relationship between item discrimination indices (a) and item difficulty indices of the individuals chosen according to their ID number's being odd and even, and according to being in sub or super groups is analyzed, it is seen that there are relations in 0,01 significance level, and positively medium and high level relations. According to these findings, it can be said that item parameters invariance assertion for 2PL model is

verified. After determining the item and ability parameter invariance and appropriate model, analysis and the results for 2PL model are given from this part of the study.

Results

After the data to be analyzed is recorded in the .prn extended formatted text format, commands required for analysis is shown in Figure 4.

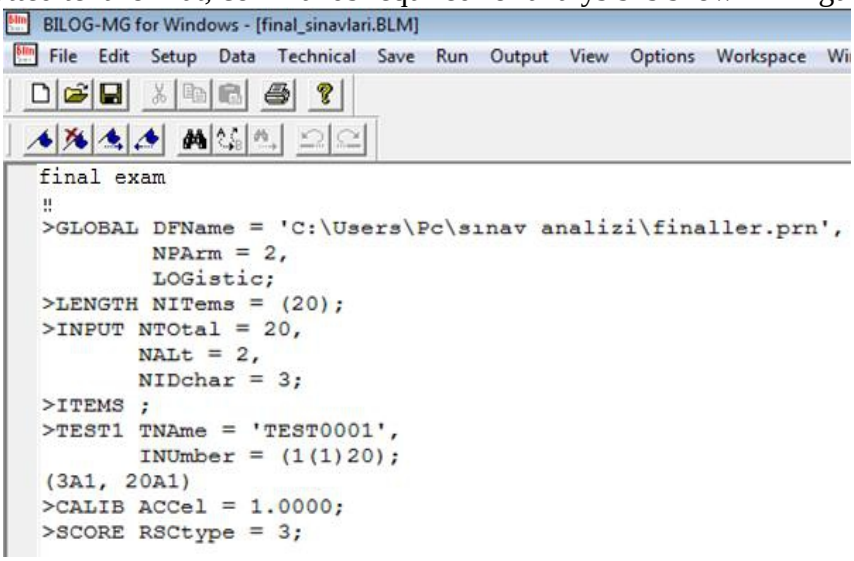

Figure 4. Commands for data analysis on 2PLM

After the data file is identified, the accuracy of commands is controlled and the analysis process is started by remarking "run". BILOG MG program gives three different kinds of outputs. The first of these is the Phase 1 output in which descriptive information related to items is located.

Results for Phase 1 and Comments

In this phase, how each item in the given table is processed by the program, how many people answered the related item, how many of the answerers responded this item correctly, the percentage of the correct response to the item, logic value and biserial correlation value that show the correlation between the item and test and known as discrimination are involved. The values acquired for the final exam that is composed of 20 items are shown in Table 10.

Table 10. Item statistics about test items for 2PLM

Item	#Tried	#Right	PCT	Logit	Item-Test Correlation	
					Pearson	Biserial
1	771.0	323.0	41.9	0.33	0.339	0.429
2	771.0	241.0	31.3	0.79	0.217	0.284
3	771.0	255.0	33.1	0.70	0.248	0.322

4	771.0	219.0	28.4	0.92	0.391	**0.520**
5	771.0	295.0	38.3	0.48	0.364	0.463
6	771.0	310.0	40.2	0.40	0.269	0.341
7	771.0	257.0	33.3	0.69	0.135	0.175
8	771.0	208.0	27.0	1.00	0.241	0.323
9	771.0	229.0	29.7	0.86	0.300	0.397
10	771.0	332.0	43.1	0.28	0.335	0.422
11	771.0	332.0	43.1	0.28	0.339	0.428
12	771.0	253.0	32.8	0.72	0.235	0.306
13	771.0	231.0	30.0	0.85	0.252	0.332
14	771.0	231.0	30.0	0.85	0.331	0.436
15	771.0	315.0	40.9	0.37	0.371	0.469
16	771.0	321.0	41.6	0.34	0.281	0.355
17	771.0	254.0	32.9	0.71	0.239	0.310
18	771.0	238.0	30.9	0.82	0.202	0.265
19	771.0	234.0	30.4	0.83	0.356	0.469
20	771.0	321.0	41.6	0.34	0.315	0.399

When Table 10 is analyzed, it is seen that the easiest item is 10 and 11th items that have the percentage of the same respondence (43,1 %). Logit value that belongs to both items is calculated as 0,28. Besides this, as for the most difficult one of the items in the test is 8th item with 27 % percentage of respondence and 1,00 logit values. When the items in the test are compared according to their discrimination indices, it has been identified that the most discriminating question is the 4th question. Furthermore, when the discrimination values of the other items are analyzed, it can be said that the other items except 2, 7 and 18th items can be used as they are in the test without making correction or by making little corrections (Atılgan, Kan ve Doğan, 2013). The first part of the analysis is completed with this table.

Results for Phase 2 and Commends

In the second part of the analysis, there are quadrat points for EM and Newton cycles, gradient values calculated related to the items, at which iteration the convergence is completed, ranges set for chi-square calculations and theta ability values for these ranges. Offset and theta values determined for chi-square values calculated related to the items are given in Table 11.

Table 11. Information about test items for 2PLM

Interval Counts for Computation of Item Chi-squares							
11.	63.	179.	211.	132.	65.	39.	71.
Interval Average Thetas							
-1.684	-1.185	-0.741	-0.278	0.190	0.713	1.184	2.351

When Table 11 is analyzed, how many people are located in the ranges set for chi-square calculations and theta ability level threshold values related to them are given. It is seen that the minimum ability required for chi-square calculation related to the items -1,684 and as for the maximum ability level is 2,351. After chi-square threshold values are determined, the values in which item parameters related to the 20 items in the test are ranked are given in Table 12.

Table 12. Estimated parameters for test items on 2PLM

Item	Intercept S.E	Slope S.E	Threshold S.E	Loading S.E	Asymptote S.E	Chısq (Prob)	Df.
1	-0.369	0.991	0.372	0.724	0.00	22.4	0.7
	0.083*	0.125*	0.096*	0.888*	0.00*	(0.0021)	
2	-0.848	0.608	1.394	0.520	0.00	15.2	0.7
	0.083*	0.097*	0.241*	0.083*	0.00*	(0.0334)	
3	-0.766	0.661	1.159	0.551	0.00	6.7	0.7
	0.082*	0.096*	0.196*	0.880*	0.00*	(0.4612)	
4	-1.136	1.162	0.977	0.758	0.00	6.4	0.7
	0.098*	0.141*	0.122*	0.092*	0.00*	(0.3820)	
5	-0.562	1.091	0.515	0.737	0.00	15.4	0.7
	0.086*	0.133*	0.097*	0.090*	0.00*	(0.0311)	
6	-0.429	0.689	0.623	0.567	0.00	5.3	0.7
	0.078*	0.102*	0.147*	0.084*	0.00*	(0.6243)	
7	-0.716	0.386	1.857	0.360	0.00	9.1	0.7
	0.078*	0.076*	0.407*	0.071*	0.00*	(0.2435)	
8	-1.074	0.614	1.748	0.523	0.00	4.6	0.7
	0.087*	0.096*	0.281*	0.082*	0.00*	(0.7127)	
9	-0.970	0.810	1.197	0.630	0.00	7.1	0.7
	0.088*	0.105*	0.170*	0.081*	0.00*	(0.4154)	
10	-0.309	0.937	0.330	0.684	0.00	18.4	0.7
	0.081*	0.121*	0.097*	0.088*	0.00*	(0.0104)	
11	-0312	1.024	0.305	0.715	0.00	16.2	0.7
	0.083*	0.125*	0.090*	0.087*	0.00*	(0.0235)	
12	-0.777	0.647	1.200	0.543	0.00	9.8	0.7
	0.082*	0.096*	0.203*	0.080*	0.00*	(0.2020)	
13	-0.936	0.720	1.301	0.584	0.00	5.3	0.7
	0.086*	0.102*	0.199*	0.083*	0.00*	(0.6266)	
14	-0.971	0.881	1.102	0.661	0.00	16.3	0.7
	0.088*	0.117*	0.161*	0.088*	0.00*	(0.0227)	
15	-0.429	1.113	0.385	0.744	0.00	24.1	0.7
	0.085*	0.137*	0.087*	0.091*	0.00*	(0.0011)	
16	-0.366	0.731	0.501	0.590	0.00	6.7	0.7
	0.079*	0.105*	0.130*	0.085*	0.00*	(0.4608)	
17	-0.765	0.613	1.248	0.523	0.00	5.3	0.7
	0.081*	0.093*	0.222*	0.079*	0.00*	(0.6198)	
18	-0.852	0.517	1.648	0.459	0.00	4.7	0.7
	0.081*	0.087*	0.307*	0.077*	0.00*	(0.6923)	
19	-0.971	0.974	0.997	0.698	0.00	10.5	0.7

	0.090*	0.123*	0.139*	0.088*	0.00*	(0.1635)	
20	-0.374	0.866	0.432	0.654	0.00	3.9	0.7
	0.081*	0.115*	0.110*	0.087*	0.00*	(0.7865)	

Category threshold parameters designate the position of item characteristic slopes and it represents the ability level necessary for answering above j threshold category in 0.50 probability (Embretson and Reise, 2000; Tang, 2006). When the table is analyzed, it is seen that threshold parameters are valued between 0,305 and 1,857. According to these received values, it can be said that the range of threshold parameters is not so wide.

Embretson and Reise (2000;334) stated that it can be commented as item discrimination of slope (a) parameter related to the item parameters in 2PL model and as item difficulty of threshold parameter (b). Due to the fact that 2PL model was used in the study, prediction parameter known as the possibility of correct responding by chance (c) is calculated as zero for all items. When discrimination parameters related to the items in Table 12 are analyzed, it is determined that the 4th item is the item that has the highest discrimination and 18th item is the one with the lowest discrimination. Besides this, when item difficulties are analyzed, the item determined as the most difficult item is the 7th item, as for the easiest item is 11th item.

Results for Phase 3 and Commands

In this phase of the study, average standard deviation values of ability distributions before passing to ability predictions for all of the 771 students who attended the final exam as part of the study are given. In order to generate normal distribution, the average of ability predictions must be 0 and standard deviation must be 1 (Hambleton and Swaminathan, 1985). As the average (0.01) standard deviation (0.972) values obtained for ability distribution in the study are very close the intended value, it can be said that predicted ability distribution shows normal distribution.

BILOG MG, gives information about the reliability of the test beside ability predictions related to individuals. Reliability coefficient obtained in the study is calculated as 0,719 and it has been reached to the conclusion that final exam questions are reliable at acceptable level according to this value. Moreover, the variance amount that the test explained according to coefficient calculated about the variance of the test at related BILOG MG output is determined as 28,03 %. Item information functions that belong to items are analyzed while the possible reasons for the explained variance amount's being low are being searched. As a result of this research, especially when item information function slopes are analyzed, it is seen that the information amount that 7, 8, 17 and 18th items give is too little and it is

thought that the variance's being low that the test explained is derived from the related items' giving too little information.

The Analysis of the Relation between Ability and Error parameters Predicted with the Help of Different Programs

In the study, correlation analysis is done with the purpose of determining what kind of relation there is between ability parameters and standard error values predicted related to individuals with the help of different programs. Ability predictions of the 771 individuals in the study are realized with the help of BILOG MG, IRT PRO and JMETRİK packaged programs. The correlation analysis results are given between ability parameters predicted with the help of different programs in Table 13.

Table 13. Correlation analysis results for estimated ability parameters

Programs	BILOG MG	IRT PRO	JMETRİK
BILOG MG	1		
IRT PRO	0,9996**	1	
JMETRİK	0,9998**	0,9999**	1

According to the correlation analysis results, it is seen that there is a perfect relation between ability parameters predicted by 3 different programs. Although D invariant is taken as 1,7 in BILOG and JMETRİK programs, IRT PRO makes predictions without taking this invariant into consideration. As a result, it is seen that there is a perfect relation between ability parameters predicted related to individuals. While there is a perfect relation between predicted ability parameters, descriptive statistic results done with the purpose of determining what kind of distribution the average and standard deviation values show in ability parameters' study group is given in Table 14.

Table 14. Ability parameters average and standard deviation values predicted by different programs

	$\bar{x}$	ss	N
BILOG MC	.0027	.8392	771
J METRİK	-.0322	.8353	771
IRT PRO	-.0002	.8493	771

According to the values acquired in Table 14, it is seen that ability parameters average and standard deviation values predicted by BILOG, JMETRİK and IRT PRO are quite close to each other. Even though ability parameters average values are close to each other, it is seen that the values predicted by IRT PRO are quite close to the intended values. Furthermore in the study, analysis results done with the purpose of determining the direction

and severity of the relation between standard error parameters related to ability predictions is given in Table 15.

Table 15. Correlation analysis results for residuals

Programs	Error _BILOG MG	Error _IRT PRO	Error _JMETRİK
Error _BILOG MG	1		
Error _IRT PRO	.9633**	1	
Error _JMETRİK	.9630**	.9996**	1

The relations between error parameters predicted by 3 different programs are seen according to the correlation analysis results given in Table 15. There is a perfect relation between error parameters predicted by JMETRİK and IRTPRO programs, there is a high relation as for between BILOG program and IRTPRO and JMETRİK. It is determined that both ability and error parameters predicted with the help of different programs to this acquired results take very close values to each other.

Conclusions, Discussion and Suggestions

The following results are reached according to the findings within the scope of this study.

Statistically meaningful relations are found according to the correlation analysis results between form X and form Y groups, which consist of the items placed in the test with the aim of analyzing the invariance of ability parameters and are composed randomly, and the groups that are composed of single and double items, thus the invariance assumption of ability parameters has been provided. However, it is determined that correlation coefficient between the groups is at low level when response patterns related to the students are divided into 4 groups randomly as A, B, C, D forms and for this reason this assumption can not be met when they are divided into 4 groups as A, B, C, D.

It has been determined that there are statistically meaningful relations at medium and high level in positive direction as a result of the correlation analysis done for the groups composed as the case of single and double being of ID numbers and sub and super groups of 27% in respect to abilities in the analyses done for the invariance of item parameters, thus the invariance assumption of item parameters is assured.

In the study, it is determined that there are statistically meaningful relations at high level and in positive direction, as a result of the correlation analysis results done with the purpose of determining what kind of relation there is between the ability parameters predicted related to individuals with the help of different programs. It is determined that the ability parameters predicted with the help of BILOG MG, IRT PRO and JMETRİK packaged

programs are similar. However, the relation between IRTPRO and JMETRİK programs is higher compared to the others.

In the study, it is determined that there are statistically meaningful relations at high level and in positive direction, as a result of the correlation analysis results done with the purpose of determining what kind of relation there is between the error parameters predicted related to individuals with the help of different programs. It is determined that the error parameters predicted with the help of BILOG MG, IRT PRO and JMETRİK packaged programs are similar, the relation between IRTPRO and JMETRİK programs is higher compared to the others.

According to the results acquired in the study, the following suggestions are made for researchers and practitioners.

- The sampling of this study is composed of 771 students. A wider sampling can be used in the similar studies that will be done in the future.
- In the study, the data set related to mathematics test was used. In the studies with similar topics that will be done in the future, the data set related to the exam results of a different course can be used.
- In the study, 1PLM, 2PLM and 3PLM were applied to the two categorized data set. Except these 3 logistic models, 4PLM can also be used in the studies with similar topics that will be done in the future.
- As the data set used in the study was graded in two categories, MTK models used in the two categorized grading data were used in the study. In any of the studies that are thought to be done in the future and in which two categorized data will be used; the statistic results related to KTK and the statistic results related to MTK can be compared.
- In the study, the ability and error parameters predicted related to the individuals are determined by using different packaged programs. When there will be done studies on the same topic, comparisons from the aspect of ability and error parameters by using different computer programs can be made.

References:

1. Çakıcı Eser, D. (2013) PISA 2009 Okuma Testinden Elde Edilen İki Kategorili Verilerin Bilog Programı ile İncelenmesi, Eğitim ve Öğretim Araştırmaları Dergisi, 2 (4), 142-153.
2. Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2010). *Sosyal Bilimler İçin Çok Değişkenli İstatistik Teknikleri*. Ankara: Pegem Akademi.
3. Embretson, S. E. and Reise, S. P. *Item Response Theory for Psychologists,* New Jersey: Lawrence Erlbaum Associates Publishers.
4. Hambleton, R. K. & Swaminathan, H. (1985). Item Response Theory: Principles and Applications, Boston: Kluwer

5. Hambleton, R. K., Swaminathan, H. and Rogers, H. (1991). *Fundamentals of Item Response Theory.* Newbury Park CA: Sage.
6. Huang,Y. F., Tsou, M. Y., Chen, E. T., Chan, K. H. and Chang, K. Y. (2013) Item response analysis on an examination in anesthesiology for medical students in Taiwan: A comparison of one- and two-parameter logistic models, Journal of the Chinese Medical Association, 76 (6), 344–349
7. Güler, N., Uyanık, G. K., Teker, G. T. (2014) Comparison of classical test theory and item response theory in terms of item parameters, European Journal of Research on Education , 2, 1 - 6
8. Kalaycı, Ş. (2010). *SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri*. Ankara: Asil Yayın Dağıtım.
9. Köse, A. (2010). *Madde Tepki Kuramına Dayalı Tek Boyutlu ve Çok Boyutlu Modellerin Test Uzunluğu ve Örneklem Büyüklüğü Açısından Karşılaştırılması.* Yayımlanmamış Doktora Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.
10. Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems.* New Jersey: Lawrence Erlbaum Associates Publishers.
11. Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. New York: Addison-WesleyPub. Co.
12. McKinley, R.L. (1989). Methods, plainly speaking: An introduction to Item Response Theory. Measurement and Evaluation in Counseling and Development, 22, 37-57.
13. Meyer, J. P., (2014) Applied Measurement with Jmetrik, Paperback: 149 pages. Publisher: Routledge.
14. Nenty, H. J., Adedoyin, O. O. (2013) Test For Invariance: Inter and Intra model Validation of Classical Test and Item Response Theories, Asia Pacific Journal Research, 1 (9), 2320-5504.
15. Rup, A.A. & Zumbo, B.D. (2006). Understanding parameter invariance in unidimensional IRT models. Educational and Psychological Measurement, 66(1), 63-84.
16. Stout, W. (1987). A Nonparametric approach for assessing latent trait unidimensionality. *Psychometrica, 52.*
17. Tabachnick, B.G. and Fidell, L.S. (2001). *Computer-assisted research design and analysis.* Boston: Allyn& Bacon.
18. Tavşancıl, E. (2005). Tutumların Ölçülmesi ve SPSS ile Veri Analizi. Ankara: Nobel Yayınları.
19. Uyanık, G. K., Teker, G. T., Güler, N., An Investigation Of Goodness Of Model Data Fit: Example Of Pisa 2009 Mathematics Subtest,
20. International Journal on New Trends in Education and Their Implications, 4319 (2), 188 – 196.

21. Weiss, D. J.,& Minden, V. S. (2012) Technical Report: A Comparison of Item Parameter Estimates from Xcalibre 4.1 and Bilog-MG., Assessment Systems Corporation.
22. Zheng, X, &Rabe-Hesketh, S (2007). Estimating parameters of dichotomous and ordinal item response models using gllamm. *The Stata Journal*, *7*, 313–333.