EK 2: Üçgenler Konusu Zihin Haritası Örneği

THE FIT OF ONE-, TWO-, THREE-PARAMETER MODELS OF ITEM RESPONSE THEORY TO ÖZDEBIR ÖSS EXAM MATHEMATICS TEST

Dr. İsmail ÖNDER
Ortaöğretim Fen ve Matematik Alanları Eğitimi

ABSTRACT

The present study aimed to investigate whether one of the IRT models fit the data obtained from ÖZDEBİR ÖSS Exam mathematics subtest and distinguish the IRT model that fits well to the data. Data was derived randomly from examinees throughout Ankara (N=1097). Goodness of fit investigations were done through examination of unidimensionality, local independence, equal discrimination indices, minimal guessing, non speeded test administration, invariance of ability parameter estimates and invariance of item parameter estimates. In addition, item information functions and item characteristics curves were reviewed. Results presented that the most appropriate model data fit was achieved by 2-PLM.

Keywords: item response theory, model data fit analysis, person and item statistics

MADDE TEPKİ KURAMINA DAYANAN BİR-, İKİ-, ÜÇ-PARAMETRELİ MODELLERİN ÖZDEBİR ÖSS SINAVI MATEMATİK TESTİNE UYUMU

ÖZET

Bu araştırma, ÖZDEBİR ÖSS Sınavının matematik testinden elde edilen veriye Madde Tepki Kuramına dayanan modellerden birinin uyup uymadığını ve bu veriye en uygun modeli belirlemeyi amaçlamıştır. Veri, sınava Ankarada giren öğrencilerden rasgele seçilerek elde edilmiştir (N=1097). Model veri uyumu araştırmaları tek boyutluluk, yerel bağımsızlık, eşit ayırtedicilik gücü, minimum şansla doğru cevaplandırma, hızlandırılmamış test uygulaması, yetenek parametresi kestirimlerinin değişmezliği ve madde parametreleri kestirimlerinin değişmezliğinin soruşturulması ile yapılmıştır. Ayrıca, madde bilgi fonksiyonları ve madde karakteristik eğrileri gözden geçirilmiştir. Elde edilen sonuçlar, iki parametreli modelle en iyi model veri uyumunun elde edildiğini göstermiştir.

Anahtar Sözcükler: madde tepki kuramı, model veri uyumu analizleri, kişi ve madde istatistikleri

1. INTRODUCTION

Item Response Theory (IRT) consists of family of probabilistic models that hypothesize the relationship between an examinee's latent ability and a correct response to an item (Cohen, Bottge & Wells, 2001; Li, Lissitz & Yang, 1999; Stone & Hansen, 2000). The three most popular IRT models are one-parameter logistic model (1-PLM), two-parameter logistic model (2-PLM) and three-parameter logistic model (3-PLM). These models are appropriate for dichotomously scored data and differ in the number of parameters used to describe items (Hambleton et al., 1991). The most complex one of these models is 3-PLM. In this model, item difficulty parameter, item discrimination parameter, and item guessing parameters are estimated. When the guessing parameter is assumed to be zero, the 3-PLM is reduced to the 2-PLM for which only item difficulty parameter and item discrimination parameter needed to be estimated. A further restriction can create the 1-PLM. In that model, item discrimination parameter is treated as if all items have equal and fixed item discrimination parameters. Therefore, in that model, only item difficulty parameter is estimated (Kelkar, Wightman & Luecht, 2000; MacDonald & Paunonen, 2002).

All IRT models include a set of assumptions about the data to which the model is applied. Unidimensionality assumption is the most important assumption common for all IRT models. According to unidimensionality assumption only one ability is measured by a set of items in a test (Cohen, Bottge & Wells, 2001; Hambleton et al., 1991). Local independence is another assumption which means that examinees responses to any pair of items are uncorrelated when the ability influencing test performance is hold constant (Hambleton et al., 1991). Another assumption is non-speeded test administration which implies that all examinees must attempt to answer all items of the test. 1-PLM and 2-PLM have another assumption called minimal guessing and 1-PLM have additional assumption called equal discrimination indices. Violations in these assumptions may result in erroneous IRT model parameter estimates (Fan, 1998). Therefore, the choice of the model depends on the verification of model assumptions.

When a given IRT model fits the test data of interest, several desirable features are obtained. Ability estimates obtained from different sets of items will be the same and item parameter estimates obtained in different groups of examinees will be the same. Moreover, several researchers presented the benefits of IRT models in testing applications (Camilli & Shepard, 1994; Fan, 1998; Hambleton & Swaminathan, 1985; Stone & Hansen, 2000). However, the advantages of IRT models can only be obtained when fit exist between model and the test data. When the model fits the data set of interest, invariant item and person statistics will be obtained. Therefore, in goodness of fit analysis after the examination of model assumptions, invariance property of item and person statistics should be investigated. In other words, first model assumptions should be checked and then expected model features should be checked. Comparison of the fits of different

models to the data set will facilitate the choice of an appropriate model (Hambleton et al., 1991).

Although there are many studies related to IRT measurement framework, there are not many studies conducted that investigate goodness of model fit.

1.1. Purpose of the Study

This study focused on the fit of different models of Item Response Theory (IRT) to mathematics subtest data of the ÖZDEBİR ÖSS D-II Exam. Therefore, the main purpose of this study was to investigate whether the mathematics test data fit one of the IRT models and by which IRT model the best fit was achieved.

2. METHOD

2.1. Data

The data set was obtained from the ÖZDEBİR ÖSS D-II Exam that was applied nationwide in 2004. The sample was selected randomly from examinees that took the test throughout Ankara. The sample was composed of 1097 examinees that were either in their final year of high school education or already graduated a high school. Therefore, the sample was composed of students whose ages were ranging from 17 to 20. 529 of students were female while the remaining 567 of them were male students.

2.2. Instrument

The ÖZDEBİR ÖSS 2004 D-II Exam is an achievement test. This exam emphasis high school curriculum which consist of 180 multiple choice items in a single form. There are four subtest in the form and each has 45 items. In other words, there are 45 items related to Turkish, 45 items related to social sciences, 45 items related to mathematics items and 45 items related to science. In this study, mathematics subtest was examined. Examinees' mathematics test performance is presented in Table 2.1.

Table 2.1 Mathematics Test Performance (N=1097)

Tests	Mean	Median	SD
Mathematics	29.0	31.0	10.1

2.3. Procedure

In this study the test used has been designed for students who was preparing for Student Selection Test (ÖSS) in Turkey. The sample was selected randomly from the examinees throughout Ankara. Then, four subsamples were formed from the data set such as Gender, Ability, Odd-Even and Difficult-Easy. The Gender sample was composed of two subsamples; male and female. Female subsample was

composed of 529 female students and Male subsample was composed of 567 male students. Ability subsample was composed of to subsamples; Low and High. Examinees whose total test performance fall within the 0^{th} and 60^{th} percentile formed the Low ability subsample and the remaining examinees formed High ability sample. Both Gender and Ability subsamples were used to examine invariance property of item parameter estimates. Mathematics test performance of Gender and Ability subtests were presented in Table 2.2. Odd-Even subsample was composed of two subsamples; Odd and Even. All the odd items in mathematics subtest formed Odd subsample and all the even items of mathematics subtest formed Even subsample. Difficult-Easy subsample was composed of two subsamples; Difficult and Easy. Mathematics Items that have p-values greater than 0.6 formed Easy subsample and the other remaining items in mathematics subtest formed Difficult subsample. Both Odd-Even and Difficult-Easy subsamples were used to investigate invariance property of ability parameter estimates. Descriptive statistics of Odd-Even and Difficult-Easy subsamples were presented in Table 2.3. ITEMAN from the Assessment of the Micro CAT^{tm}, Testing System were used to determine the item difficulties (p-value) and item discrimination indices (biserial and point biserial values). SPSS 11.5 for Windows was used to determine descriptive statistics and BILOGMG 3.0 for Windows was used while predicting item and person statistics. Finally, goodness of fit investigations was done and Item Characteristic Curves (ICC) and Item Information Functions (IIF) obtained by the help of BILOGMG program were investigated to determine best informative items.

Table 2.2 Mathematics Test Performance of Gender and Ability Samples

Group	N	Mean	Median	SD
Gender				
Female	529	27.0	28.0	9.99
Male	567	30.9	33.0	9.76
Ability				
High ability	450	38.1	38.0	3.93
Low ability	647	22.7	24.0	8.05

Table 2.3 Descriptive Statistics of Odd-Even and Difficult-Easy Samples

Group	N	Mean p-value	Mean Biserial	Mean Point-Biserial
Odd-Even				
Odd	23	0.655	0.683	0.487
Even	22	0.633	0.719	0.534
Difficult-Easy				
Difficult	19	0.458	0.650	0.512
Easy	26	0.780	0.737	0.508

2.4. Goodness of Fit Analysis

Goodness of fit investigations were done as discussed by Hambleton et al. (1991) under two headings which are checking model assumptions and checking expected model features. In the first part, unidimensionality, local independence, equal discrimination indices, minimal guessing and non speeded test administration were investigated. In the second part, invariance of item parameter estimates and invariance of ability parameter estimates were investigated.

Unidimensionality assumption was checked through conducting principle component analysis. Eigenvalues obtained under each factor was investigated to decide on unidimensionality assumption. Local independence assumption was checked through investigating inter-item correlation matrices for whole, high ability and low ability groups. In order to check equal discrimination indices assumption, biserial and point biserial values obtained by the help of ITEMAN program were investigated. In order to check if the data meets minimal guessing assumption for the one-, two-, three-parameter models, the performance of low ability examinees on the most difficult mathematics items was reviewed. Non speeded test administration assumption was checked through investigating omitted responses toward the end of the test.

In order to investigate the degree to which the property of invariance held for the item difficulty and item discrimination parameter estimates, item parameters estimated were compared across one subsample of the sample such as male versus female. In order to investigate the degree to which the property of invariance held for the ability parameter "θ" estimates, ability parameters estimates obtained were correlated across one subsample of the sample such as odd versus even.

3. RESULTS

3.1. Checking Model Assumptions

Unidimensionality

In order to examine whether the unidimensionality assumption was met in mathematics test, factor analysis was conducted and the scree plot was obtained. The presence of a dominant first factor was treated as an evidence for unidimensionality (Hambleton et al., 1991). The eigenvalues of first three factors are presented in Table 3.1. The eigenvalue of a first factor is 3.5 times greater than the second factor. Moreover, the first factor accounts for 27.5% of the total variance; therefore, the first factor seems to be dominant. Moreover, the scree plot obtained also presented the existence of a single dominant first factor. As a result, it can be concluded that the unidimensionality assumption is met.

Table 3.1 Eigenvalues of First Three Factor

	First Factor	Second Factor	Third Factor
Eigenvalues	12.37	3.62	1.58

Local Independence

In order to investigate local independence, the inter-item correlations of whole, low ability and high ability groups were examined. The mean value of inter item correlations of high and low ability groups are close to zero and lower than the value obtained for whole group (see Table 3.2). This indicates that the local independence assumption is met.

Table 3.2 Inter-Item Correlations of Whole, High Ability and Low Ability Groups.

	Groups		
	Whole	High Ability	Low Ability
Mean	0.245	0.048	0.144

Equal Discrimination Indices

The variability of item biserial and point-biserial values obtained by ITEMAN program was used in decisions made on the degree of violation of the equal discrimination assumption. The variances of biserial and point biserial values of mathematics items were 0.022 and 0.013 which are close to zero indicating that both discrimination parameters do not vary a lot (see Table 3.3). This indicates that the equal discrimination assumption is met.

Table 3.3 Descriptive Statistics of Biserial and Point-Biserial Values

Discrimination Parameters	N	Mean	SD	Variance
Biserial	45	0.700	0.149	0.022
Point-biserial	45	0.510	0.112	0.013

Minimal Guessing

The performance of low ability examinees on most difficult mathematics items was investigated. The performance of low ability examinees on some of the most difficult items was high (see Table 3.4). For example, 22.1% of the low ability examinees correctly responded to item 22 and similarly 21.0% of the low ability examinees answered correctly the item 29. In general, the performance of low ability students on most difficult items is expected to be low. However, results indicated that low ability students' performance on some difficult items was high. Therefore, it can be concluded that minimal guessing assumption was not hold.

Table 3.4 Correct Response Percentages of Low Ability Students (N=647).

Items	p-values	Frequency	Percent Correct
Item 16	0.284	120	18.5
Item 22	0.374	143	22.1
Item 23	0.139	54	8.3
Item 29	0.380	136	21.0
Item 42	0.361	78	12.1

Non-Speeded Test Administration

In order to examine whether the test was non-speeded or not, percentage of examinees completing the initial and final five items in mathematics subtest were reviewed. Percentage of examinees that did not marked the first and last 5 items are presented in Table 3.5. As seen in the table percentage of students who did not complete the last five items are high compared to students that did not complete the first five items. However, percentage of students who did no responded to first and second item is also high, 28.1 and 32.5, respectively. This strange result could be obtained because of the hardness of these items. In addition, remaining items at the beginning of the test have low missing values. Moreover, the test was applied in specific time limit. In other words, students were supposed to finish test in 180 minutes. Therefore, some of the students may not reach some of the items placed through the end of the test. The results imply that the non-speeded test administration is not viable.

Table 3.5 Omitted Response Percentages on First and Last Five Items

	First 5 Items			Last 5 Items	
	Number Missing	Percent Missing		Number Missing	Percent Missing
Item 1	308	28.1	Item 41	380	34.6
Item 2	356	32.5	Item 42	471	42.9
Item 3	10	0.9	Item 43	472	43.0
Item 4	24	2.2	Item 44	461	42.0
Item 5	26	2.4	Item 45	404	36.8

3.2. Checking Expected Model Features

In order to decide whether the mathematics test data fit one of the IRT models and by which IRT model the best fit was achieved, it is necessary to investigate invariance property of item and ability parameter estimates obtained by each IRT models. The better fit is achieved when the model of interest produces more invariant ability and item statistics.

Invariance of Item Statistics

In order to investigate the degree to which the property of invariance held for both difficulty parameter and discrimination parameter under each model, item parameters obtained on female subsample were correlated by item statistics obtained on male subsample. A similar investigation was done with item statistics obtained on high ability and low ability samples. 1-PLM was not included in investigation of invariance property of discrimination parameter since in 1-PLM fixed discrimination parameter is used.

Table 3.6 Correlations of Item Statistics Obtained on Different Samples

Invariance Across	Item Difficulty Parameter			Item Discrimination Parameter		
	1-PLM	2-PLM	3-PLM	1-PLM	2-PLM	3-PLM
Female-Male Sample	0.967**	0.969**	0.973**	NA	0.895**	0.805**
High-Low Ability Sample	0.840**	0.843**	0.889**	NA	0.545**	0.356*

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).
Note. NA= Not Applicable

The correlations obtained under difficulty parameter for all IRT models were strong indicating that by all IRT models invariance property of difficulty parameter was achieved. The highest correlations were obtained under 3-PLM and the second highest correlations were obtained under 2-PLM. Correlation of difficulty parameters obtained on Ability subsamples are quite low compared to that obtained on Gender subsamples. This indicates that as the variability between subsamples increases the correlations obtained decreases. Moreover, correlations obtained in investigations on invariance property of discrimination parameter under 2-PLM were strong in both sampling conditions. However, moderate to high correlations were obtained under 3-PLM. In other words, correlations obtained under 3-PLM were lower than that obtained under 2-PLM. In addition, as in results obtained in difficulty parameter, correlations obtained in investigations on invariance property of discrimination parameter decreases as the variability between subsamples increases. In addition, it is observed that correlations obtained under item difficulty parameter were higher compared to correlations obtained under item discrimination parameter.

Invariance of Ability Parameter Estimates

In order to investigate the degree to which the property of invariance held for the ability parameter estimated under each model, person statistics obtained on odd subsample were correlated with person statistics obtained on even subsample. Similarly, person statistics obtained on difficult subsample were correlated by person statistics obtained on easy subsample.

Table 3.7 Correlations of Person Statistics Obtained on Different Samples

Invariance Across	IRT Models		
	1-PLM	2-PLM	3-PLM
Odd-Even Sample	0.776**	0.771**	0.766**
Difficult-Easy Sample	0.862**	0.837**	0.428**

** Correlation is significant at the 0.01 level (2-tailed).

Correlations obtained in investigations on invariance property of ability parameter estimates under each model were strong except the correlation obtained on Difficult-Easy subsample under 3-PLM. In other words, more invariant ability parameters were estimated by 1-PLM and 2-PLM compared to 3-PLM. In general, results indicated that person and item statistics obtained were invariant.

3.3. Graphical Fit Plots

ICCs and IIFs of each item obtained by each IRT model were investigated to decide which model fits the data better. Fit judgments provided that the overall best fit was achieved by 2-PLM. Fit judgments made on each item indicated that 42.2% (n=19) of the time the best fit to test data was obtained under 2-PLM, 35.5% (n=16) of the time the best fit to test data was obtained under 3-PLM and finally 22.2% (n=10) of the time the best fit to test data was obtained under 1-PLM.

4. CONCLUSION

The present study examined whether the mathematics test data fit one of the IRT models and by which IRT model the best overall fit to data could be achieved. Goodness of model data fit investigations was conducted. Therefore, model assumptions and expected model features were checked. In these investigations 1097 examinees' data on mathematics test were used. The data was selected randomly throughout participants in Ankara. Results of these analysis presented how the assumptions and features of each model reacted to test data.

It is important to investigate to what extent the IRT model assumptions are valid for the given data and how well IRT model fits the data since violation of IRT model assumptions may lead to erroneous IRT model parameter estimates (Fan, 1998). Investigations done on model assumptions presented that unidimensionality, local independence and equal discrimination indices assumptions were hold. However, questionable results were obtained while investigating minimal guessing assumption since low ability students responded highly in some difficult items. In other words, although in general low ability students performed poorly in difficult items, there were difficult items that were answered correctly by majority of low ability students. Therefore, results of this investigation did not support the assumption of minimal guessing. In addition, non-speeded test administration assumption was also not hold. Percentage of students that did not respond the questions toward the end of the test was high. In general forty percent of the

examinees did not respond to last five mathematics items. On the other hand majority of students answered the items placed at the beginning of the test. However, first two items of mathematics test were also not answered by majority of students. The percent correct values of these two items are quite low indicating the hardness of these items which then may result in such high missing values. In addition, the test was an achievement test which was administered in a specific time limit. Therefore, students had a limited time to complete the entire test. Therefore, some of the students because of time limit could not reach to the end of the test.

Important model-data fit information can be obtained by investigating the property of invariance (Leeson & Fletcher, 2003). Hambleton, Swaminathan and Rogers (1991) indicated that, when IRT model fits the data, parameters that characterize an item do not depend on examinees' ability distribution and the parameters that characterize an examinee do not depend on the sets of test items. Therefore, high correlations obtained for both item statistics and person statistics under each IRT model can be considered as an evidence for existence of property of invariance. Results of this investigation presented that assumption of invariance was hold for difficulty parameter estimates under each IRT model since high correlations were obtained under each IRT model. By contrast more invariant discrimination parameters were obtained under 2-PLM compared to 3-PLM since under 2-PLM all correlations were high however under 3-PLM moderate to high correlations were observed (see Table 3.6). Correlations obtained in investigations on invariance property of ability parameter estimates under 1-PLM and 2-PLM was high in all sampling conditions. However, low correlations were observed under 3-PLM while correlating ability parameters estimated on Difficult subsample and Easy subsample. Investigations conducted on ICCs and IIFs indicated that the overall best fit to the data was achieved by 2-PLM. Therefore, although minimal guessing assumption showed that the guessing did occur among low ability students, analysis in general presented that 2-PLM have almost perfect fit to data. In other words, 2-PLM provided the most appropriate fit for the test data.

REFERENCES

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items* (vol. 4). Thousand Oaks, CA: Sage.

Cohen, A.S., Bottge, B.A., & Wells, C.S. (2001). "Using Item Response Theory to assess effects of mathematics instruction in special populations". *Council for Exceptional Children*, 68 (1), 23-44.

Fan, X. (1998). "Item response theory and classical test theory: An empirical comparison of their item/person statistics". *Educational and Psychological Measurement*, 58 (3), 357-381.

Hambleton, R.K., Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Kelkar, V., Wightman, L.F., & Luecht, R.M. (2000, April). "Evaluation of the IRT parameter invariance property for the MCAT". *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, New Orleans, LA.

Leeson, H., & Fletcher, R. (2003, December). "An investigation of fit: Comparison of 1-, 2-, 3- parameter IRT models to project asTTle data". *Paper presented at the Joint NZARE/AARE Conference*, Auckland.

Li, Y. H., Lissitz, R.W., & Yang, Y.N. (1999, April). "Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items". *Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec*, Canada.

MacDonald, P. & Paunonen, S. V. (2002)." A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory". *Educational and Psychological Measurement*, 62 (6), 921-943.

Stone, C.A. & Hansen, M.A. (2000). "The effect of errors in estimating ability on goodness-of-fit tests for IRT models". *Educational and Psychological Measurement*, 60 (6), 974-991.

İsmail ÖNDER
Dr., Ortaöğretim Fen ve Matematik Alanları Eğitimi
e115251@hotmail.com