International Journal on New Trends in Education and Their Implications
July 2013 Volume: 4 Issue: 3 Article: 19 ISSN 1309-6249

AN INVESTIGATION OF GOODNESS OF MODEL DATA FIT: EXAMPLE OF PISA 2009 MATHEMATICS SUBTEST

Res. Assist. Gülden KAYA UYANIK
Sakarya University
Education Faculty
Department of Measurement and Evaluation in Education
Sakarya, TURKEY

Inst. Gülşen TAŞDELEN TEKER
Sakarya University, Education Faculty
Department of Measurement and Evaluation in Education
Sakarya, TURKEY

Assist. Prof. Dr. Neşe GÜLER
Sakarya University, Education Faculty
Department of Measurement and Evaluation in Education
Sakarya, TURKEY

ABSTRACT

Although Classical Test Theory (CTT) has been used for test development, Item Response Theory (IRT) is beginning to major theoretical source. However, the model-data fit should be verified as a prerequisite. Therefore, in this study it is aimed to investigate which IRT model will provide the best fit to the data obtained from PISA 2009 mathematics subtest. For goodness-of-fit analysis, first the model assumptions and then the expected model features were tested. The model assumptions unidimensionality, local independence and non-speeded test administration were investigated. In the expected model features part the invariance of ability parameter estimates and invariance of item parameter estimates were analyzed. In addition, item characteristics curves and item information functions were analyzed. To determine the best model, two different ways were followed: first number of items which fits with the model and then the results of the ki-square statistics of -2 log likelihood values of models were compared. The results suggested that two parameter logistic model is the most appropriate model for data fit.

Key Words: Item response theory, model data fit analysis, person and item statistics.

INTRODUCTION

One of the advantages of Item Response Theory (IRT) compared to Classical Test Theory (CTT) is that item and test parameters can be predicted independently of the group and the group members' properties. The advantages of IRT is true only when the model-data fit is achieved at a satisfactory level. When the concordance is low, invariance of item and ability parameters cannot be attained. The model-data fit is achieved by satisfying the primary assumptions. The assumptions necessary for all IRT models are: assuring unidimensionality and local independence, and making sure that the test is not a speed test (Hambleton, Swaminathan and Rogers, 1991; Baker, 2001; Embretson and Reise, 2000). This study examines firstly the model assumptions and then the invariance of the predictions for item and ability parameters- the expected model properties.

One of the suppositions of IRT is "unidimensionality". What is meant by the term unidimensionality is the measurement of one single ability with one test, and it is too difficult to meet this assumption exactly. In order

to achieve this, it is sufficient to have a dominant factor affecting the test performance. The dominant factor represents the ability to be measured with the test (Hambleton and Swaminathan, 1985). Unidimensionality means measurement of only one variable with the items. In other words, it means that there is only one property leading the answering behaviour for the items. However, unidimensionality is an assumption too difficult to confirm in practice. Besides, when there is one dimension in unidimensionality studies, it might not always be easy to predict what the dimension is and to describe it (Crocker and Algina, 1986). For this, another dimension leading the answering behaviour, a factor, is mentioned. Unidimensionality work is generally done with factor analysis. Additionally, correlations used in local independence studies can also be employed (Lord 1980; Hambleton and Ravinelli 1986; Crocker and Algina 1986; Harvey and Hammer 1999; quoted by Doğan, 2002).

An assumption equivalent to the assumption of unidimensionality is local independence. Local independence points to the statistical independence of the responses given by sub-groups of a certain ability level to an item. For this assumption to be correct, an individual's performance in an item should not influence his/her performance in the other items. That is to say, it means that the correct or incorrect answer given to an item is independent of the correct or incorrect answer given to another item (Doğan, 2002). An item, for instance, should not give clues as to answering the other items, or the answer given to an item should not be a prerequisite to answering the other items. This can be represented with the equation given below (Hambleton and Swaminathan, 1985):

$$f\left(\frac{x}{\theta}\right) = \sum_{ui=x} \prod_{i=1}^{n} P_i(\theta)^{ui} Q_i(\theta)^{1-ui}$$

Another assumption is whether or not the test is a speed test. Hambleton, Swaminathan and Rogers (1991) contend that one of the criteria used to determine whether a test functions like a speed test is that 75% of the group should answer 80% of the test. This same criterion was used in this study so as to find whether the matmenatics test in the research was a speed test.

The expected model properties of the Item Response Theory is the invariance of the predictions for item and ability parameters. One of the major properties of all item response models is that item parameters are invariant; that is to say, they are independent of the group. Hambleton and Swaminathan (1985) pointed out that the invaiance of item parameters could be checked by comparing the item parameters obtained from two or more sub-groups (such as boys and girls, top group, bottom group) taken from the population.

Another property of item response models is the invariance of predictions for ability parameters. Hampleton and Swaminathan (1985) claimed that the invariance of ability parameter predictions could be analysed by comparing the predictions made by selecting two or more samples from the relevant pool of samples. In this research, the comparability of ability parameters obtained from items of different difficulty levels for the same group of students was checked for the purposes of analysing the invariance of ability predictions. For our purposes, the items were classified as easy-difficult and even-odd; and Pearson's Moments Multiplication Correlations between ability predictions obtained separately for each individual from easy-dificult and even-odd items according to 1-, 2-, and 3-parameter logistic models were examined.

This research aims to analyse the 2009 PISA mathematics test scores in 1-, 2-, and 3-parameter logictic models, and to determine the most appropriate model for item scores. In line with this purpose, firstly model assumptions were examined separately, and then the invariance of ability parameter predictions- which are the expected model proerties- was examined.

METHOD

Research Model

This is a descriptive study analysing model-data fit for PISA 2009 mathematics sub-test in 1-, 2-, and 3-parameter logistic models according to item response theory.

Data Source

The research data were composed of the scores for 12 items belonging to 1127 people who had taken the 2009 PISA mathematics sub-test in Turkey. The item scores for the individuals were obtained by encoding 1 for correct answers and 0 for incorrect answers. The descriptive statistics for 2009 PISA mathematics sub-test are shown in Table 1.

Table 1. Descriptive Statistics

N	1127
K	12
Ortalama	4,88
Ortanca	5
Tepe Değer	4
Std. Sapma	2,524
Varyans	6,370
Çarpıklık Katsayısı	0,298
Basıklık Katsayısı	-0,518

N: number of person, K: number of ıtem

As is clear from Table 1, the average for 2009 PISA mathematics sub-test is smaller than the median, and the skewness coefficient receives a positive value close to zero. Based on these findings, it may be said that the scores display almost normal distribution. Based on skewness coefficient's receiving a negative value, mathematics test scores may be said to be more skew than the normal (Büyüköztürk, Köklü, and Çokluk, 2012).

Analyses of Data

This research firstly examines the model assumptions and then analyses the expected model properties. The assumptions of unidimensionality, local independence and unaccelerated test application were investigated in the model assumptions part while the invariance of ability parameter predictions and the invariance of item parameter predictions were investigated in the model properties part. In addition to that, item characteristic curves and item information functions were also analysed. Two ways were pursued in deciding on the model with the best model-data fit. Firstly, the number of items fitting the model was decided on. As a second way, the chi-square results were compared with -2 likelihood values obtained from the models. The SPSS, STATISTICA, ITEMAN and BILOG programmes were used for the analyses.

FINDINGS

Unidimensionality

In order to test whether or not the PISA mathematics test met the unidimensionality assumption, factor analysis was performed. Because the item scores obtained from the test had two categories, factor analysis was done through tetrakoric correlation matrix by using the STATISTICA programme. The eigenvalues for the factor analysis and the findings concerning the explained variance percentages are shown in Table 2.

Table 2: Factor Analyses Results of PISA 2009 Mathematics Subtest

Factor	Eigenvalues	% of Variance	Cumulative %
1	4,843	40,36	40,36
2	1,111	9,26	49,62

40.36% of the variance in the PISA 2009 mathematics sub-test is accounted for by the first factor. Besides, the fact that there is a sharp decrease in the proportions between eigenvalues and accounting for the variance after the first factor and that the difference between them becomes fourfold is remarkable. Based on these results, the PISA 2009 mathematics test may be said to be unidimensional, or to measure only one structure.

In order to be able to identify unidimensionality visually, the factor eigenvalue chart shown in Figure 1 was examined.

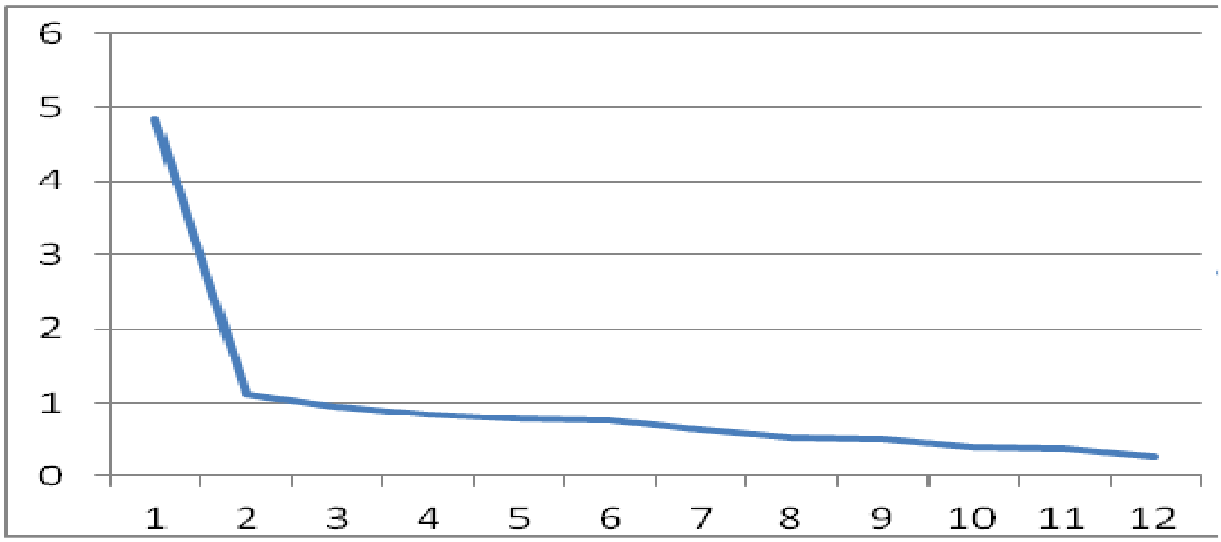


Figure 1: Factor Eigenvalue Chart

According to Figure 1, the slant forms a plateau after the second point; which means that the contibutions made by factors after this point are small and almost the same. For this reason, it is thought that the number of factors available is one. In this case, it might be said that the scale is unidimensional.

Local Independence

If the items are not statistically independent for constant level of ability, the test scores of some individuals are expected to be higher than those of other individuals. In consequence, it will be possible to account for the individual's test performance with more than one ability. The violation of this assumption also means the violation of unidimensionality assumption. Hence, explanation of the relations in a set of items should be connected to only one ability (Hambleton and Swaminathan, 1985). Thus, the 12 items in the PISA 2009 mathematics sub-test achieving the assumption of unidimensionality were also regarded as achieving the assumption of local independence.

Non-Speeded Test Administration

Following the analysis, it was found that 85% (967 people) of the group of 1127 people answered the whole test whereas 89% (1006 people) answered 75% of the test. With this result, the 75% criterian set by Hambleton, Swaminathan and Rogers (1991) was attained. That the majority of the group had answered the whole test was regarded as the indicator of the fact that the test was not a speed test. Moreover, the answering rate for the final item in the test was analysed, and the answering status of the individuals for the item was shown in Table 3.

Table 3: Answering Rate For The Final Item

	Frequency	%
Whole	1127	100
Answered	1102	97,8
Skipped	25	2,2
Correct	162	14,7
Incorrect	940	85,3

According to Table 3, 1102 out of 1127 people (97.8%) answered the last item of the test. 25 people ((2.2%) skipped the item while 162 people (14.7%) answered it correctly, and 940 (85.3%) answered it incorrectly. Thus, it was found that more than ¾ of the group had answered the item. On examining the answers given to the whole test by those who had skipped the item, it was observed that they did not answer the items also in the other items beginning with item 1. In this case it might be interpreted as that they skipped the item not because they could not reach the item due to time restrictions, but because they did not have the behaviour tested by the item. Thus, it was established that the test did not function as a speed test.

Invariance of Item Statistics

This study analysed the item parameters predicted under each model through correlations between item parameters that were predicted by forming different groups. The groups were divided in two different ways. First, the group of 1127 was ordered according to total scores, and thus it was divided into two according to individuals receiving high and low scores. Individuals receiving low scores were those whose score was in the 0%-60% range, and individuals rereceiving high scores were those whose score was in the 60%-100% range. In selecting these intervals, the criterion required that the individuals had answered **a minimum of** 2/3 of the test for the 60%-100% range, and they had answered **a maximum of** 2/3 of the test for the 60%-0% range. Second, the group of 1127 people was divided into two on the basis of gender as girls and boys.

Item parameters specific to 1-, 2-, and 3- PL models were predicted for all groups. Because the number of items was smaller than 30, the invariance of item parameters were analysed by comparing Spearman's non-parametric rank differences Correlation Coefficients (rho) calculated between item parameters that were predicted in different groups for each model. The correlations between item parameters predicted for each group are shown in Table 4.

Table 4: Invariance of Item Statistics: Correlations of Item Statistics Obtained on Different Samples

	Parametreler					
	Item-Difficulty (b)			Item-discrimination (a)		Guessing (c)
	1 PL	2PL	3PL	2PL	3PL	3PL
High-low ability	1,00**	0,998**	0,993**	0,991**	0,974**	0,991**
Female-male	0,986**	0,988**	0,980**	0,644**	0,554**	0,778**

**p<0,01

According to Table 4, it may be said that there is a perfect correlation between item difficulties predicted according to 1-PL model (r=1.00, p<0.01), and there are very high and significant correlations between item difficulties predicted according to 2-PL model (r=0.998, p<0.01) and between item difficulties predicted according to 3-PL model (r=0.993). A similar case is also observed on dividing the group according to gender. The strongest correlation forms in the girls and boys groups between item difficulty values predicted according to 2-PL model; but there are also high and significant correlations between item difficulty parameters predicted according to 1-PL (r=0.986) and 3-PL models (r=0.980). The high and significant correlations observed between item difficulty values that were obtained from different groups indicate that the invariance of item difficulty is achieved for all models. Besides, the correlation values obtained from gender groups with low and high scores decreased by 0.14 in the 1PL model, 0.10 in the 2PL model, and 0.13 in the 3PL model. Based on these findings, the interpretation might be that the best invariance of item difficulty parameter is obtained in the 2PL model.

The invariance of item distinctiveness parameters which were obtained in the 2PL and 3PL models was analysed in groups with low and high scores and in gender groups. Although high and significant correlations were found between the predicted distinctiveness parameters in 2PL (r=0.991, p<0.01) and 3PL models (r=0.974, p<0.01) in the low score and high score groups, the correlation value obtained in the 2PL model is bigger. In gender groups, a higher correlation was obtained between distinctiveness parameters in the 2PL model (r=0.644, p<0.01), but a medium level significant correlation was observed in both models. In the light of these findings, it may be said that the invariance between distinctiveness parameters is not as strong as that of item difficulty parameter, but that it is achieved best in the 2PL model.

And lastly, correct responses by chance predicted from the low and high score groups according to 3PL model- that is to say, the invariance of c parameter- was analysed. According to Table 3, there are significant and very high correlations (r=0.991) between c parameters predicted from low score and high score groups, and significant and high correlations (r=0.778, p<0.01) between c parameters predicted from boys and girls groups. Based on these findings, the c parameter may be said to achieve independence from the group.

Invariance of Ability Parameter Estimates

In analysing the invariance of ability predictions, the comparability of ability parameters obtained for students of the same goup from items with different levels of difficulty was checked. For that purpose, the items were classified as easy-difficult, and as even-odd numbered, and Pearson's Moments Multiplication Correlations between ability predictions obtained separately for each individual from easy-dificult and even-odd numbered items according to 1-, 2-, and 3-parameter logistic models were examined.

In doing the easy-difficult classification for the items, the items below average difficulty were regarded as difficult, and those above the average difficulty were regarded as easy.

Table 5: Item Difficulties of PISA 2009 Mathematics Sub-Test

Number of Items	Item difficulties
1	0,433
2	0,404
3	0,326
4	0,654
5	0,055
6	0,546
7	0,170
8	0,522
9	0,871
10	0,253
11	0,506
12	0,144

According to Table 5, items 4,6, 8, 9, and 11 are the easy items; and 1, 2, 3, 5, 7, 10, and 12 are the difficult ones. The average for item difficulty indices was 0.26 for the difficult items in the mathematics test whereas the average was 0.62 for the easy items. The average item difficulty index for items with odd number was 0.40 while it was 0.42 for the items with even number. Table 6 shows the correlations between ability predictions obtained from easy-difficult and even-odd numbered items for each model.

Table 6: Invariance of Ability Parameter: Correlations of θ Values Obtained on Different Samples

	1-PL Model	2-PL Model	3-PL Model
Easy-diff.	0,823**	0,820**	0,819**
Odd-even	0,790**	0,796**	0,778**

**p<0.01

International Journal on New Trends in Education and Their Implications
July 2013 Volume: 4 Issue: 3 Article: 19 ISSN 1309-6249

On examining the correlations in Table 6, it may be said that there is a positive high level correlation between scores obtained from easy-difficult and even-odd numbered items and ability parameters predicted with 1-,2-, and 3-parameter logistic models. The correlation expressing the invariance between ability predictions obtained from the easy and difficult items whose item difficulty averages were different increased by 0.033 in the 1 PL model, by 0.024 in the 2 PL model, and by 0.041 in the 3 PL model, compared to the correlation expressing the invariance between ability predictions obtained from even and odd-numbered items with very close item difficulty averages. On comparing the models, it might be sid that the invariance of ability predictions assumption is best achieved in the 2 PL model.

The fit between PISA 2009 mathematics test and 1 PL, 2 PL, and 3 PL- which are the models of item response theory- was examined through the BILOG programme. Two ways were followed in analysing the model-data fit. In the first one, the numbers of items fitting the model were compared. In the other method, the results of chi-square statistics performed were compared with -2 loglikelihood values obtained from the models. Table 7 shows the values compared in relation to the model-data fit.

Table : Values compared in relation to the model-data fit

	1PL	2PL	3PL
-2loglikelihood	14522,2967	14389,0691	14406,0247
Number of fitting ıtems	6	8	5

The item concordance statistics obtained through analyses were examined for each model. In deciding the item concordance of the items, the chi-square statistics were taken into consideration. Thus, in the research it was concluded that items yielding a p value bigger than 0.05 fitted the model. On examining the values, it was found that only 6 of the 12 items in the test fitted the 1-parameter model whereas 8 items fitted the 2-parameter model, and 5 fitted the 3-parameter model. Based on these findings, it was concluded that the 2-parameter model was more congruous in terms of model-data fit.

Another method suggested by Emretson and Reise (2000) for the evaluation of model congruity is the difference between -2 log likelihood values. The -2 log likelihood value of the data shows the degree of model's differentiation from the data. The betterment of 2 PL model for PISA 2009 mathematics test compared to 1 PL model can be evaluated with this calculation:

$$X^2 = (-2\log likelihood_{1PL}) - (-2loglikelihood_{2PL})$$
$$= 14522,2967 - 14389,0691$$
$$= 133,2276$$

The degree of freedom for X^2 is the number of increasing parameters in a more complex model (for the 2 PL model). Because there were 12 items in PISA 2009 mathematics test, compared to the 1 PL model, an increase of 12 item distinctiveness parameter occurred in the 2 PL model. Hence, the degree of freedom is 12. The value for 12- freedom degree is 21.0267. Because..... , the hypothesis is refuted. Based on this result, the interpretation may be that the 2 PL model causes a significant difference in terms of the betterment of the model. In a similar vein, it may be calculated that the 3 PL model's bettering according to the 2 PL model is not significant.

$$X^2 = (-2\log likelihood_{3PL}) - (-2loglikelihood_{2PL})$$
$$= 14406,0247 - 14389,0691$$
$$= 16,9556$$

The degree of freedom for X^2 is the number of increasing parameters for 3 PL model, which is a more complex model this time.

Because there were 12 items in PISA 2009 mathematics test, compared to the 2 PL model, an increase of 12 chance parameter occurred in the 3 PL model. Hence, the degree of freedom is 12. since the value for 12-

freedom degree is 21.0267, the hypothesis is accepted. Based on this result, it may be said that the 2 PL model does not cause a significant difference in terms of the betterment of the model.

In addition to the results obtained, considering the fact that the assumption of the invariance of item and ability parameters prediction is achieved best in the 2 PL model, it may be concluded that the best fitting model for PISA 2009 mathematics test is the *2-parameter logistic model.*

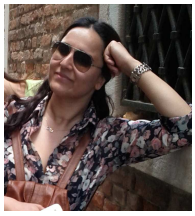
RESULTS AND DISCUSSION

This research aimed to analyse the model-data fit holding for the items in PISA 2009 mathematics sub-test. In the scope of concordance statistics, firstly the assumptions of item response theory were focussed on. It was found that the data satisfied the assumptions of unidimensionality, local independence and that the test is not a speed test. In the second step, the invariance of item and ability parameters was analysed.

Within the scope of the research, 2009 PISA mathematics sub-test was evaluated with 1-, 2-, and-parameter logistic models; and an attempt was made to determine the best fitting model for the item scores. The goodness of fit results showed that the best fit was achieved with the 2 PL model. In line with these findings, it may be stated that the studies and analyses to be performed in relation to PISA 2009 mathematics sub-test based on item response theory would yield more accurate results with the 2 PL model.

IJONTE's Note: This article was presented at 4th International Conference on New Trends in Education and Their Implications - ICONTE, 25-27 April, 2013, Antalya-Turkey and was selected for publication for Volume 4 Number 3 of IJONTE 2013 by IJONTE Scientific Committee.

BIODATA AND CONTACT ADDRESSES OF AUTHORS



Gülden KAYA UYANIK is a research assistant of the Measurement and Evaluation Department in Faculty of Education, Sakarya University, Sakarya-Turkey. She is a graduate student at the Hacettepe University, working towards her PhD in Measurement and Evaluation in Education. Her research interest is measurement and evaluation in education; generalizability theory, statistics and research methods in social sciences. She has authored a book about generalizability theory, co-authored or presented many articles, and conference presentations.

Res. Asst. Gülden KAYA UYANIK
Sakarya University, Education Faculty
Department of Measurement and Evaluation
54300 Hendek/Sakarya-Turkey
E. Mail: guldenk@sakarya.edu.tr

Gülşen TAŞDELEN TEKER is an instructor at the Sakarya University, Turkey. She worked as a science and technology teacher for 2 years. After that she has begun to work as an instructor since three years. Gülşen is a graduate student at the Hacettepe University, working towards her PhD in Measurement and Evaluation in Education. Her research interest is generalizability theory, standard setting, differential item functioning, and statistical analysis. She has co-authored a book called Generalizability Theory and has presented many articles and conference presentations.

Gülşen TAŞDELEN TEKER
Sakarya University, Faculty of Education
Hendek, Sakarya, Turkey
E. Mail: gtasdelen@sakarya.edu.tr

Neşe GÜLER is assistant professor and division head of the Measurement and Evaluation Department in Faculty of Education, Sakarya University, Sakarya-Turkey. Her research interest is measurement and evaluation in education; generalizability theory, statistics and research methods in social sciences.Dr. Güler has authored two books and authored, co-authored, or presented almost fourty articles, and conference presentations.

Assist. Prof. Dr. Neşe GÜLER
Sakarya University, Education Faculty
Department of Measurement and Evaluation
54300 Hendek/Sakarya-Turkey
E. Mail: gnguler@gmail.com

REFERENCES

Baker, F. B. (2001). *The Basics of Item Response Theory* (2th Edition). USA, Heinemann.

Büyüköztürk, Ş., Köklü, N. ve Çokluk, Ö. (2012). Sosyal Bilimler için İstatistik. PegemA Yayıncılık, Ankara.

Crocker, L. ve Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York: Holt, Rınehart and Wınston

Doğan, N. (2002). *Klasik Test Teorisi ve Örtük Özellikler Kuramının Örneklemler Bağlamında Karşılaştırılması*. Yayımlanmamış Doktora Tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Embretson, S. E. ve Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates Publishers, Mahwah, New Jersey.

Hambleton, R. K. ve Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Kluwer Nijhoff Publishing, Baston.

Hambelton, R.K., Swaminathan, H. ve Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications.

Yen, W. M. ve Fitzpatrick, A. R. (2006). Item Response Theory. R. L. Brennan (Ed.) *Educational Measurement* (pp.112-153). USA, American Council on Education and Praeger Publishers.