Hacettepe Üniversitesi Eğitim Fakültesi Dergisi 32 [2007] 210-220

AN INVESTIGATION OF GOODNESS OF MODEL DATA FIT

MODEL VERİ UYUMUNUN ARAŞTIRILMASI

İsmail ÖNDER[•]

ABSTRACT: IRT models' advantages can only be realized when the model fits the data set of interest. Therefore, this study aimed to investigate which IRT model will provide the best fit to the data obtained from ÖZDEBİR ÖSS 2004 D-II Exam Science Test. In goodness-of-fit analysis, first the model assumptions and then the expected model features were checked. In the model assumption part unidimensionality, local independence, equal discrimination indices, minimal guessing, and non-speeded test administration was 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 (ICC) and item information functions (IIF) were analyzed. The results suggested that the most appropriate model data fit was achieved by two parameter logistic model.

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

ÖZET: Madde Tepki Kuramına (MTK) dayanan modellerin avantajı, model veri uyumu sağlandığında görülebilir. Bu nedenle bu çalışma ÖZDEBİR ÖSS 2004 D-II Sınavının Fen Testinden elde edilen verilere MTK'ya dayanan modellerden hangisinin en iyi uyum sağlayacağını araştırmayı amaçlamıştır. Model veri uyumu çalışmalarında öncelikle model sayıltıları ve daha sonra beklendik model özellikleri incelenmiştir. Model sayıltıları kısmında tek boyutluluk, yerel bağımsızlık, eşit ayırtedicilik gücü, minimum şansla doğru cevaplandırma ve hızlandırılmamış test uygulaması sayıltıları incelenmiştir. Model özellikleri kısmında ise yetenek parametresi kestirimlerinin değişmezliği ve madde parametreleri kestirimlerinin değişmezliği incelenmiştir. Ayrıca, madde karakteristik eğrileri ve madde bilgi fonksiyonları incelenmiştir. 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

Two popular measurement frameworks for interpreting test scores are Item Response Theory (IRT) and Classical Test Theory (CTT). CTT was used over the majority of 20th century (Demirtaşlı, 2002; Traub, 1997) and are still used (Bechger et al., 2003). However, several researchers presented some shortcomings of CTT (Camilli & Shepard, 1994; Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Mellenbergh, 1996). In addition, it was found ineffective in solving some measurement problems such as equating of test scores, identification of biased items, linking and building item banks (Demirtaşlı, 2002; Fan, 1998). On the other hand, all the limitations presented are overcome by the use of IRT. IRT is a measurement approach that relates the probability of a particular response on an item to overall examinee ability (Camilli & Shepard, 1994). Therefore, in IRT ability parameters estimated are not test dependent and item statistics estimated are not group dependent. Some of the advantages of IRT over CTT were presented by several researchers (Camilli & Shepard, 1994; Fan, 1998; Hambleton et al., 1991; Özdemir, 2004). However, measurement specialists can not benefit from these advantages unless model data fit is achieved (Fan, 1998; Hambleton et al., 1991). Although there are several studies that investigated model data fit (Leeson & Fletcher, 2003; Yalçın, 1999), few studies investigated the fit of IRT models to data obtained from achievement tests administered to high school students. Therefore, in this study model data fit investigations were conducted on the data obtained from examinees that were preparing for Student Selection Test (SST) in Turkey.

[•] Dr, Ortaöğretim Fen ve Matematik Alanları Eğitimi, e115251@metu.edu.tr

Model data fit investigations can be applied under two sections (Hambleton et al., 1991) which are checking model assumptions and checking expected model features. The first part investigates whether the test data satisfies the assumptions of the particular model of interest. In second part, property of invariance obtained by each model for both person and item statistics is investigated. In addition, goodness of model data fit can be investigated through analysis of Item Characteristics Curves (ICC) and Item Information Functions (IIF).

IRT framework includes a group of models where applicability of each model depends on the nature of the test items and the viability of different theoretical assumptions about the test items (Fan, 1998). For test items that are dichotomously scored there are three IRT models which are one parameter logistic model (1-PLM), two parameter logistic model (2-PLM) and three parameter logistic model (3-PLM). Fit analysis is conducted for dichotomous data under these models. A primary distinction among these models is the number of parameters used to describe items. In 1-PLM only the item difficulty “b” parameter is estimated. In addition, fixed discrimination “a” parameter is used and no pseudo change “c” parameter is estimated. In 2-PLM both item difficulty and item discrimination parameters are estimated. However, as in 1-PLM, in 2-PLM no guessing parameter is estimated. In other words, 2-PLM makes no allowance for guessing behavior. On the other hand 3-PLM estimates item discrimination, item difficulty and pseudo chance parameters.

1.1. Purpose of the Study

The main purpose of this study is to investigate whether the ÖZDEBİR ÖSS 2004 D-II Exam science subtest data fit one of the IRT models.

2. METHOD

2.1. Data Source

The data set was obtained from examinees that took the ÖZDEBİR ÖSS 2004 D-II Exam in 2004. This exam is applied nationwide and all the ÖZDEBİR Dershane’s which are the private institutions founded to help student in preparing for SST in Turkey administered this test to its students. The examinees were selected randomly from Dershane’s throughout Ankara. The selected sample was composed of 1097 examinees. The age range of the sample was 17 to 20. The sample was composed of examinees which are in their final year of high school education and examinees that are already graduated a high school.

2.2. Data Collection Instruments

The ÖSS 2004 D-II Exam is an achievement test which was consisted of 45 mathematics and 45 science items as well as 90 items related to Turkish and social sciences. The science part contains 19 physics, 14 chemistry and 12 biology related items. Although there were items related to physics, chemistry and biology, items in each part in general were designed to assess students’ science performance. Therefore, in the following sections science part as a whole will be treated as a subtest that was constructed conceptually to assess students’ science performance, of the exam. The test emphasizes high school curriculum and 180 minutes were given in order to complete the test. The test was administered under standard conditions all over the country. Below the performance characteristics of respondents are given.

Table 2.1 Performance Characteristics of Respondents on Science Test (N=1097)

Tests	Mean	Median	SD	Skewness	Kurtosis	Minimum	Maximum
Science	25.2	25.0	9.62	-0.012	-0.914	3	45

2.3. Sampling

Several samples were formed from the data set discussed above in order to test IRT model assumptions, invariance and fit plots. While forming First 20-Last 20, Odd-Even and Difficult-Easy samples, science test as a whole was treated as a single sub sample since items in the science test conceptualy was developed to assess students’ science performance.

Gender Sample

Sample of female participants and those of male participants were selected from the data set. Science subtests performance of both female and male examinees was investigated to examine invariance property of item parameters obtained by each IRT model. The performance of male and female examinees on science subtest is presented in Table 2.2.

Ability Sample

The whole sampling group was sorted according to test scores and then low and high ability groups were formed. Low ability sample was formed from examinees whose scores fall within the 0th and 60th percentile. High ability sample was formed from examinees whose scores fall within the 60th and 100th percentile range. These percentile ranges were choosen since examinees between 60th and 100th percentile ranges at least corrcltly answered 2/3 percent of the items and examinees between 0th and 60th percentile range at most answered 2/3 of the items correctly. The performance characteristics of high and low ability samples were presented in Table 2.2. Science subtests performance of both high ability and low ability examinees was investigated to examine invariance property of item parameters obtained by each IRT model and test local independence and minimal guessing assumptions of IRT models.

Table 2.2 Performance Characteristics of Gender and Ability Samples on Science Test

Group	N	Mean	Median	SD	Skewness	Kurtosis
Gender						
Female	529	23.1	22.0	9.12	0.220	-0.803
Male	567	27.1	28.0	9.65	-0.252	-0.801
Ability						
High ability	450	34.5	34.0	4.86	-0.043	-0.393
Low ability	647	18.7	19.0	6.16	-0.121	-0.570

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

Group	N	Mean p-value	Mean Biserial	Mean Point-Biserial
First-Last 20				
First 20	20	0.547	0.633	0.489
Last 20	20	0.525	0.597	0.452
Odd-Even				
Odd	22	0.557	0.638	0.481
Even	23	0.560	0.588	0.447
Difficult-Easy				
Difficult	18	0.370	0.589	0.455
Easy	27	0.685	0.628	0.469

First 20-Last 20 Sample

Sample of initial 20 items and last 20 items on science test was selected. In other words, first twenty questions in science test formed first 20 sample and last twenty questions formed the last 20 sample. The descriptive statistics of first 20 and last 20 samples were presented in Table 2.3. Ability parameters estimated on these samples were investigated to examine invariance property of ability parameters estimated by each IRT models.

Odd-Even Sample

Sample of odd items and even items were selected from science test. In other words, all the odd items in the test were included to odd sample and similarly, all the even items were included to even sample. The descriptive statistics of both odd and even samples were given in Table 2.3. Ability parameters estimated on these samples were investigated to examine invariance property of ability parameters estimated by each IRT models.

Difficult-Easy Sample

Sample of difficult items and easy items were selected from science test. In other words, p-vales of items were investigated and items which have p-values greater than 0.5 were placed to easy sample and the other remaining items were placed in difficult sample. The descriptive statistics of both difficult and easy samples were given in Table 2.3. Ability parameters estimated on these samples were investigated to examine invariance property of ability parameters estimated by each IRT models.

2.4. Design of the Study

Methods for assessing goodness of fit were presented by Hambleton et al. (1991). Goodness of fit investigations were done under two headings which are checking model assumptions and checking expected model features. In the first part degree to which model assumptions held was investigated through analysis of unidimensionality, local independence, equal discrimination indices, minimal guessing and non speeded test administration. In the second section, degree to which desired model features were obtained was investigated through analysis of invariance of item parameter estimates and invariance of ability parameter estimates. Further investigations were done on the number of misfitting items identified by each model. Moreover, ICC and IIF obtained by each IRT model for each item were investigated deeply.

2.5. Preliminary Analysis

Dichotomously scored data was used while estimating IRT parameters. Therefore, before starting goodness of fit studies, each item was investigated deeply in order to see whether there were any flawed items. The results indicated that although there were some items which have alternatives that should be revised, the science items in general were working well and discriminating well. Therefore, no item was decided to be excluded from the study.

2.6. Goodness of Fit Analysis

2.6.1. Checking Model Assumptions

Unidimensionality

To check the unidimensionality assumption factor analysis was conducted. Eigenvalues and scree plot obtained was investigated in order to determine whether there was a dominant first factor. According to Hambleton et al. (1991) a dominant first factor is needed to satisfy unidimensionality assumption. In other words, there should be a large difference between the first eigenvalue and second eigenvalue. Moreover, a significant drop in the contribution of the factors between the first and second factors can be seen as an evidence for unidimensionality. Reckase (1979) recommended that the first factor should account for at least twenty percent of the variance in order to obtain reasonable ability estimates and stable item parameters.

Local Independence

To check the local independence, inter-item correlations for whole group and for two subgroups (high and low ability groups) were obtained. The mean values of inter-item correlations of sub-groups were expected to be close to zero and lower than the whole group which indicates that the local independence assumption is met. According to Hambleton et al. (1991) when unidimensionality assumption is met the local independence assumption is also satisfied. On the other hand, while unidimensionality assumption is satisfied, if the items in the test have some other dimensions such as clue for detecting the correct answer for some examinees or pre-requested skills other than performance being measured, local independence may not hold. Therefore, science test was reviewed to investigate whether there are such items.

Equal Discrimination Indices

To check the equal discrimination indices assumption the distribution of biserial and point-biserial values obtained from ITEMAN analysis were analyzed. Homogeneous distribution of biserial and point-biserial values can be considered as an evidence for satisfying equal discrimination indices assumption (Hambleton et al., 1991). Point-biserial values are bias corrected therefore they were also included in the analysis. In other words, point-biserial values are bias corrected since the contribution of an item score to the total score was removed before calculating the item discrimination parameter.

Non-Speeded Test Administration

To check non-speeded test administration assumption, number of omitted responses toward the end of the test was investigated. In other words, it is investigated whether the majority of students were able to reach at the end of the test in a given time limit. In addition, percentage of examinees that did not respond to first 5 and last 5 items was reviewed. If similar percentages are obtained, non-speeded test administration assumption can be assumed to be met.

Minimal Guessing

To check the minimal guessing assumption, the performance of low ability examinees on most difficult items was investigated. The difficult items were selected by the help of p-values obtained from ITEMAN results. The items with p-values which are lower than 0.3, were selected as difficult. The performance of low ability examinees on most difficult items should be low. If it happens the minimal guessing assumption can be assumed to be met.

2.6.2. Checking Expected Model Features

Invariance of Item Statistics

To investigate the degree to which the property of invariance held for the item difficulty and item discrimination parameter estimates, item parameters estimated by all IRT models under different sampling strategies (male vs. female or high ability vs. low ability samples) for each item were correlated. High correlation was treated as an evidence for invariance of item statistics. In addition to correlation analysis, the scatter plots were investigated in order to check the strength of the relationship. However, one parameter logistic model was not included into the invariance of discrimination parameter analysis since this model assumes fixed discrimination parameter; therefore, it is not possible to conduct correlation analysis.

Invariance of Ability Parameter Estimates

To investigate the degree to which the property of invariance held for the ability parameter “θ” estimates, ability parameters estimated on difficult vs. easy, first 20 vs. last 20 or odd vs. even samples were correlated. In addition to correlation analysis, scatter plots were investigated in order to check the strength of the relationship. If high correlations are obtained between θ values, the invariance of ability parameter estimates can be treated as hold.

2.7. Graphical Fit Plots

ICCs obtained by BILOGMG for the 1-PLM, 2-PLM and 3-PLM were examined. In other words, in order to compare observed and predicted test score distribution, BILOGMG was used to generate fit plots for each item under each model. The fit plots obtained were investigated to determine which models observed test score distribution provides closest fit to the predicted test distribution.

3. RESULTS

3.1. Checking Model Assumptions

Unidimensionality

Results presented that the data showed four factors that did not meet strictly with assumption of unidimensionality (see Table 3.1). However, the test was assumed to be unidimensional since according to Hambleton et al. (1991) if there is a large difference between first factor eigenvalue and second largest then unidimensionality assumption is considered as meet. In other words, unidimensionality was not strictly provided; however, the test was assumed to be unidimentional since the first factor was dominant and a rapid falling from first factor’s eigenvalue to second factor’s eigenvalue was observed. The ratio between first factor’s eigenvalue and second factor’s eigenvalue was 5.01 which indicates that there is a large difference between first and second factors’ eigenvalues. Moreover, results of principal axis factoring indicated that 20, 10, 8, 7 items were loaded under first, second, third and fourth factors, respectively. In addition, amount of variance explained by the first factor was also investigated considering Reckase’s (1979) recommendation. The first factor explains

22% of the total variance; however, the second largest factor just accounts for 4% of the total variance (see Table 3.1). Therefore, unidimentionality assumption was assumed to hold.

Table 3.1 Total Variance Explained

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	9.914	22.031	22.031
2	1.944	4.320	26.351
3	1.722	3.826	30.177
4	1.281	2.847	33.024

Local Independence

To check the local independence, inter item correlations for whole group and for high ability and low ability group examinees were investigated. 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. This indicates that the local independence assumption is met (see Table 3.2). In addition, all the science items were investigated deeply in order to check whether any item could provide a clue for detecting the correct answer for some examinees and whether prerequisite skills other than the performance being measured are required. The investigation showed that some of the items were problematic since these items were designed in a way that students could eliminate some of the alternatives without necessary knowledge. Therefore, although inter item correlation analysis presented that the local independence was hold, the presence of such problematic items questions the decision made on local independence assumption.

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

Groups	Mean	Minimum	Maximum	Range	Variance
Whole Group	0.198	-0.005	0.463	0.468	0.004
High Ability Group	0.054	-0.138	0.252	0.390	0.004
Low Ability Group	0.069	-0.216	0.442	0.658	0.005

Equal Discrimination Indices

To check the equal discrimination indices assumption the distribution of biserial and point biserial values obtained by ITEMAN analysis were checked. The distribution of biserial and point-biserial values had a range of 0.419 to 0.741 (Mean= 0.612, SD=0.085) and 0.329 to 0.591 (Mean= 0.464, SD=0.070), respectively. This initial finding indicates that items have non-equal discrimination indices. Table 3.3 reveals that the spreads of corresponding biserial and point-biserial values were found to be similar across all p-value intervals. Moreover, items with high p-values (0.70 and higher) according to Leeson and Fletcher (2003) may inflate high point biserial correlations. However, Table3.3 presents that items with high p-values did not inflate high discrimination values. In other words, variability was not the result of any items with high p-values in the data set.

Table 3.3 Intervals of P-Values and Corresponding Biserial and Point-Biserial Values

P-Values Intervals	Biserial Values Intervals	Point-Biserial Values Intervals
0.20-0.30	0.531-0.711	0.395-0.536
0.30-0.40	0.419-0.660	0.329-0.516
0.40-0.50	0.424-0.741	0.337-0.591
0.50-0.60	0.455-0.737	0.361-0.587
0.60-0.70	0.623-0.734	0.479-0.570
0.70-0.80	0.552-0.678	0.419-0.502
0.80-0.90	0.525-0.736	0.348-0.491

Minimal Guessing

To check minimal guessing assumption the performance of low ability examinees on most difficult items was investigated (see Table 3.4). The performance of low ability examinees on most of the difficult items was close to zero as expected. Although the performance of low ability examinees on the item 48 was quite high the minimal guessing assumption is considered to be met since examinees were reminded that the number of incorrect responses given will effect their total score calculation, before the administration of the test.

Table 3.4 Percentage of Correct Responses Given on Most Difficult Items by Low Ability Students (N=647).

Items	p-values	Frequency	Percent Correct
Item 48	0.270	97	15.0
Item 51	0.253	66	10.2
Item 63	0.289	66	10.2
Item 87	0.253	71	11.0
Item 88	0.216	59	9.1

Non-Speeded Test Administration

To check non-speeded test administration assumption number of omitted responses toward the end of the test was investigated. In addition, omitted responses on first 5 and last 5 items in science part were also investigated. It was observed that there were students that did not respond to items toward the end of the test. In addition, the response pattern of examinees on first 5 and last 5 science items was presented in Table 3.5. There is a difference between the number of omitted responses on first 5 items (9.5%) and those on last 5 items (18.8 %). However, the percentage of missing in the item 90 was quite low which indicates that most of the examinees were responded to that question although the percentages of missing in other items toward the end of the test were quite high compared to missing in first five items. This strange result questions the decision made on non-speeded test administration. However, examinees were given 180 minutes to complete the test composed of 180 items. Therefore, there will probably be some students that were not able to reach to the end of the test. Therefore, it can be concluded that the non speeded test administration assumption is not viable.

Table 3.5 Percentage of Omitted Responses on First and Last Five Items

	First 5 Items			Last 5 Items	
	Number Missing	Percent Missing		Number Missing	Percent Missing
Item 46	19	1.7	Item 86	114	10.4
Item 47	87	7.9	Item 87	203	18.5
Item 48	54	4.9	Item 88	371	33.8
Item 49	171	15.6	Item 89	275	25.1
Item 50	191	17.4	Item 90	66	6.0

So far, model assumptions were investigated. These investigations are necessary for IRT analysis. However, one can not conclude which IRT model fits better the data set from investigation done on model assumptions. Further investigation and analysis is needed to compare and decide by which IRT model best fit is achieved. Therefore, 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. In addition, number of misfitting items identified by each model was also determined and IIFs and ICCs produced were investigated while deciding which model fits the data set of interest better. These investigations were done in the following section of the study.

3.2. Checking Expected Model Features

Invariance of Item Statistics

To investigate the degree to which the property of invariance held for both difficulty parameter "b" and discrimination parameter "a" under each model, correlation analysis was conducted on samples (gender and ability) obtained by different sampling strategies (see Table 3.6). In addition, scatter plots were investigated. As discussed before, 1-PLM was not included in investigation of invariance property of discrimination parameter.

Table 3.6 Invariance of Item Statistics: Correlations of Item Statistics Obtained on Different Samples

	Item Difficulty Parameter			Item Discrimination Parameter		
Invariance Across	1-PLM	2-PLM	3-PLM	1-PLM	2-PLM	3-PLM
Female-Male Sample	0.953^{**}	0.945^{**}	0.942^{**}	NA	0.767^{**}	0.671^{**}
High-Low Ability Sample	0.933^{**}	0.869^{**}	0.895^{**}	NA	0.439^{**}	0.423^{*}

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

* Correlation is significant at the 0.05 level (2-tailed).

Note. NA= Not Applicable

All plots obtained from gender sample for the “b” parameter showed high consistency with tight convergence around the total fit line. Accordingly, correlation coefficients obtained under each model were also very strong. These strong correlation coefficients make the baseline plots under each model excellent examples of invariance. Similar results were obtained in investigations on invariance property of “b” parameter on high and low ability sample. The correlation coefficients obtained from high-low ability sample under each IRT models were also very strong. These strong correlation coefficients are excellent examples of invariance. Compared to other IRT models, correlations under 1-PLM was quite strong; therefore invariance property was best achieved under 1-PLM. Invariance of discrimination parameter was also investigated on ability and gender samples. Correlations under 2-PLM was quite strong compared to 3-PLM. In addition, correlations obtained for invariance property of discrimination parameter were weak compared to correlations obtained for item difficulty parameter. Moreover, as the variability in sample increased the correlation coefficients obtained for invariance property of both item difficulty and item discrimination parameters decreased. Compared to 3-PLM, 2-PLM provided better fit when invariance property of discrimination parameter is considered.

Invariance of Ability Parameter Estimates

To investigate the degree to which the property of invariance held for the ability parameter estimated under each model person statistics obtained on samples (difficult-easy, first-last 20 and odd-even samples) was correlated and scatter plots of correlations were also investigated. The scatter plots obtained for each IRT models indicated that the data was less scattered under 2-PLM compared to other two models. For all the IRT models the correlations of ability estimates on each sample were quite strong (see Table 3.7). 2-PLM and 1-PLM both presented comparable correlation coefficients. However, correlation coefficients obtained under 3-PLM were weak compared to other two IRT models. Therefore, according to ability parameter estimates results the 2-PLM produced more invariant ability parameter estimates.

Table 3.7 Invariance of Ability Parameter: Correlations of θ Values Obtained on Different Samples

	IRT Models		
Invariance Across	1-PLM	2-PLM	3-PLM
First 20-Last 20 Sample	0.716^{**}	0.725^{**}	0.630^{**}
Odd-Even Sample	0.846^{**}	0.852^{**}	0.831^{**}
Difficult-Easy Sample	0.777^{**}	0.776^{**}	0.723^{**}

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

In general, results presented that person and item statistics obtained from different sampling conditions were invariant.

3.3. Misfit Analysis

Number of misfitting items identified by each model provides information related to how well the observed score fits the theoretically expected score. Number of misfitting items identified by each model is presented in Table 3.8

Table 3.8 Number of Misfitting Items Identified for the Science Test (α = 0.05)

		IRT Models		
Test	N	1-PLM	2-PLM	3-PLM
Science	45	18 (40.0%)	4 (8.9%)	3 (6.7%)

The results presented that the data fit the 2-PLM and 3-PLM well since using the subject sample size of 1097, statistical test only identified 4 and 3 items as misfitting the 2-PLM and 3-PLM, respectively. The fit of the data for 1-PLM however is very questionable since 40% of the items identified as misfitting the IRT model. Hambleton et al. (1991) and Fan (1998) indicated that the consequences of such misfit are not entirely clear, therefore; results related to 1-PLM should be viewed with extreme caution.

In order to determine which model fits the data better, the -2log likelihood values obtained by each model were also investigated. The difference between -2log likelihood values obtained under each IRT model was calculated and compared with values obtained from the chi-squire table with appropriate degrees of freedom. If the obtained difference is large compared to chi-squire value, the model with smaller -2log likelihood value fits the data better. Upon looking in the chi-square table with appropriate degrees of freedom at the 95% percentiles, the critical values were obtained. The difference between -2log likelihood values and the critical value obtained were compared. Results revealed that 3-PLM fits the data better at 0.05 level of significance.

3.4. Graphical Fit Plots

In order to decide which model fits the data better, ICCs and IIFs of each item obtained by each IRT model were investigated. 3-PLM was judged to provide the overall best fit to the test data. Since, 3-PLM gave better information compared to the other IRT models. Moreover, results showed that 46.7% (n=21) of the 45 model fit judgments made indicated that 3-PLM provides the best fit to test data. The second best fit is observed under 2-PLM, since 31.1% (n=14) of the time best model data fit was observed under 2-PLM. Finally, 22.2% (n=10) of the time best model data fit was observed with 1-PLM.

4. DISCUSSION AND CONCLUSION

The purpose of this study was to investigate which IRT model would provide the best fit to the items from ÖZDEBİR ÖSS 2004 D-II Exam science test through various goodness of fit analysis. In goodness of fit analysis, first of all IRT model assumptions and then expected model features were checked. In addition, before starting these investigations, items were investigated to determine whether there are flawed items. The CTT analysis results and discussions revealed that there are no flawed items. Therefore, no item was excluded from the study.

Investigations on IRT model assumptions indicated that all assumptions were met expect the non speeded test administration and equal discrimination indices. No evidence of non speeded test administration was found since the test was administered in specific time limit and the omitted response pattern of students on first and last five items differed. Consistent with Leeson and Fletcher's (2003) findings, no evidence of equal discrimination was found among the test data. The range of biserial and point biserial values showed intermediate variation and this variation was not resulted from items with high p-values. This finding suggests that using 2-PLM and 3-PLM will provide better fit since these models include discrimination parameter in their analysis. In addition, questionable results were obtained while assessing local independence and minimal guessing assumptions. While assessing local independence assumption it is observed that there were items that students could answer correctly by eliminating alternatives by the help of given information. However, investigations done on inter-item correlations presented that local independence was achieved. While assessing minimal guessing it is observed that the performances of low ability students on some difficult items was high. However, it is also observed that majority of low ability examinees' performance on difficult items was poor. Therefore, minimal guessing assumption was held. Although some questionable results were obtained while investigating IRT model assumptions, in general most of the IRT model assumptions were held. Then, invariance property of person and item statistics obtained by each IRT model was tested in this study. According to Wright (1968) invariance property is achieved to some degree when correlation coefficients obtained are 0.80 or greater. The correlations obtained under each IRT model for ability parameter and item difficulty parameters are strong. Therefore, invariance property for ability and item difficulty parameters was held. The lowest correlation was obtained under 3-PLM. This may result from model data misfit or poor item parameter estimation (Shephard et al., 1984). Invariance of discrimination parameter estimates, on the other hand, showed variation between different sampling strategies. The highest variation is observed under 3-PLM. Although correlations obtained under 3-PLM were significant at 0.05 level, moderate correlation was observed. Similarly, moderate correlation was observed under 2-PLM; however, the correlations were significant at 0.01 level of significance. Therefore, it can be concluded that more invariant item discrimination parameters are obtained under 2-PLM. In addition, compared to correlations obtained for item difficulty parameter estimates, correlations obtained under discrimination parameter were low.

Misfit analysis revealed that the best fit was achieved with 2-PLM and 3-PLM. In other words, chi-square statistics revealed that 2-PLM and 3-PLM both fit the data well. In addition, -2log likelihood investigations also indicated that the best fit was achieved with 3-PLM. Moreover, ICCs and IIFs were investigated to judge which IRT model provides a better fit. The analysis results presented that fit judgments made on 45 items indicated that n=21 (46.7%) times the 3-PLM provided the best fit. In addition, investigations presented that in general 3-PLM provided the best information. Second best fit was observed under 2-PLM.

In general, the analysis presented that 2-PLM provides the most appropriate fit to science data. Moreover, 3-PLM appeared to fit appropriately the science data. However, results also revealed that under 1-PLM inappropriate fit was observed.

REFERENCES

Bechger, T.M., Maris, G., Verstralen, H.H.F.M., & Beguin, A.A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, 27 (5), 319-334.

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

Demirtaşlı, N.Ç. (2002). A study of raven standard progressive matrices test's item measures under classical and item response models: An empirical comparison. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 35 (2), 71-79.

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.

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.

Mellenbergh, G.J.(1996). Measurement precision in test score and item response models. *Psychological Methods*, 1, 293-299.

Özdemir, D. (2004). Çoktan seçmeli testlerin klasik test teorisi ve örtük özellikler teorisine gore hesaplanan psikometrik özelliklerinin iki kategorili ve ağırlıklandırılmış puanlanması yönünden karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 26, 117-123.

Reckase, M.D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.

Shephard, L.A., Camilli, G., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics*, 9, 93-128.

Traub, R.E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and practice*, 8-14.

Wright, B.D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conferences on Testing Problems*. Princeton, NJ: Educational Testing Service.

Yalçın, M. (1999). *The fit of one-, two-, three–parameter models of item response theory to ERDD's achievement test,* Middle East Technical University, unpublished master thesis.

Extended Abstract

IRT models' advantages can only be realized when the model fits the data set of interest. Therefore, in this study it is aimed to investigate which IRT model will provide the best fit to the data obtained from ÖZDEBİR ÖSS 2004 D-II Exam Science Test. This exam is applied nationwide by Dershanes' which are private institutions founded to help students in preparing for Student Selection Test (SST) in Turkey. The examinees were selected randomly from Dershanes throughout Ankara. The sample was composed of 1097 examinees whose age were ranging between 17 to 20. In this study science part of the test was investigated and this part was composed of 45 items related to high school curriculum. Gender, Ability, First-Last 20, Odd-Even and Difficult-Easy sub samples were formed from the sample (N=1097) in order to perform goodness of fit analysis. Gender sample was composed of male and female sub samples. In other words, all male participants (N=567) formed the male sub sample, similarly all female participants (N = 529) formed the female sub sample. Ability sub sample was composed of high ability and low ability samples. Low ability sample was formed from examinees whose scores fall within the 0th and 60th percentile range. High ability sample was formed from examinees whose scores fall within the 60th and 100th percentile range. First-last 20 samples were formed by selecting initial 20 items of science test and last 20 items of a science test, respectively. Odd-Even samples were formed by selecting odd items and even items of a science test, respectively. Difficult-Easy samples were formed by investigating p-values of items. In other words, items which have p-values greater than 0.5 were placed to easy sample and other remaining items were placed to difficult sample. Goodness of fit investigations were done under two headings which are checking model assumptions and checking expected model features. In the first part, degree to which model assumptions held was investigated through analysis of unidimentionality, local independence, equal discrimination indices, minimal guessing and non-speeded test administration. Unidimensionality was checked through investigating factor analysis results. In other words, a dominant first factor is expected to satisfy unidimensionality assumption. Local independence assumption was checked through investigation of inter-item correlation results for whole group and two subgroups (low ability and high ability). Equal discrimination indices was checked by investigating the distribution of biserial and point biserial values obtained from ITEMAN analysis. In order to check non speeded test administration assumption, number of omitted responses toward the end of the test was investigated. In order to check minimal guessing assumption, the performance of low ability examinees on most difficult items was investigated. In the second section, degree to which desired model features were obtained was investigated through analysis of invariance of item parameter estimates and invariance of ability parameter estimates. To investigate the degree to which the property of invariance was held for item statistics, item statistics estimated by all IRT models under different sampling strategies for each item were correlated. Moreover, to investigate the degree to which the property of invariance held for the ability parameter estimates, ability parameter estimated on different samples for each examinee were correlated. In addition, number of misfitting items identified by each model, item characteristics curves (ICC) and item information functions (IIF) were also investigated. Investigations on IRT model assumptions indicated that all assumptions were met expect the non speeded test administration and equal discrimination indices assumptions. A dominant first factor is observed when eigenvalues obtained and scree plot were investigated. Therefore, unidimensionality assumption was hold. The mean values of inter item correlations of high and low ability groups were close to zero and lower than the value obtained for whole group. Therefore, local independence assumption was also hold. The distribution of both biserial and point biserial values showed intermediate variation indicating that equal discrimination indices assumption was not hold. The minimal guessing assumption was hold since the performance of low ability examinees on most difficult items was low. The test was administered in a specific time limit which was 180 minutes. Therefore, some of the students were not able to reach to some of the biology items that were paced at the end of the test. Therefore, non speeded test administration assumption was not hold. Investigation on invariance property of person and item statistics obtained by each IRT model indicated that invariance property for ability and item difficulty parameters was hold since the correlations obtained were strong for both parameters. The lowest correlations for both parameters were obtained under 3-PLM compared to 2-PLM and 1-PLM. Invariance property for discrimination parameter was better achieved under 2-PLM compared to 3-PLM since correlations obtained under 2-PLM were high compared to correlations obtained under 3-PLM. In addition, correlations obtained under discrimination parameter were low compared to correlations obtained under item difficulty parameter in all IRT models. Misfit investigations indicated that the test items fit best to 2-PLM and 3-PLM. Moreover, fit judgments made by investigating ICCs and IIFs indicated that 3-PLM provided the best fit. Second best fit was observed under 2-PLM. In general, considering all results the analysis presented that 2-PLM provides the most appropriate fit to science data. 3-PLM appeared to fit appropriately to the science data; however, inappropriate fit was observed with 1-PLM.