A Study on Individualized Tests[1]

Metin YAŞAR[2]

ABSTRACT

This study aims to compare KR-20 reliability levels of “Paper and Pencil Test” developed according to Classical Test Theory and “Individualized Test” developed according to Item Response Theory (Two-Parameter Logistic Model), and the correlation levels of skill measurements obtained via these two methods in a group of students. Individualized test developed in accordance with the Two-Parameter Logistic Model was applied by means of a question pool consisting of 61 multiple-choice items which can be answered in 13 steps. On the other hand, a paper and pencil test of 47 multiple-choice items was applied to the sample student group. After the test developed according to these two methods was applied to the same group, KR-20 reliability coefficient was calculated as 0.67 for the individualized test and as 0.75 for the paper and pencil test prepared according to Classical test theory. Calculated KR-20 reliability coefficients obtained from the study were converted into Fisher Z and tested at the significance level of 0.05. No meaningful difference was detected at the 0.05 significant difference level between the two KR-20 reliability coefficients obtained from the two methods. Pearson Product-Moment Correlation Coefficient was calculated as 0.36 between the points of the individualized test and the measurement results of the paper and pencil test. A positive yet low correlation was observed between the measurement results obtained from the tests developed according to both methods. Consequently, it was seen that at the 0.05 significance level there was no statistically significant difference between KR-20 reliability coefficients of the tests developed according to the two methods and that there was a low correlation between the skill measurements of the students in both tests, but there was no significant correlation at the 0.05 significance level between the skill measurements obtained from both tests.

Key Words: Classical test theory, Individualized tests, Two-parameter logistic model, Reliability

[1] 1This study was adapted from the PhD dissertation supervised by Prof. Dr. Yasar BAYKUL.

[2] Asst. Prof. Dr. - Pamukkale University, Faculty of Education - myasar@pau.edu.tr

INTRODUCTION

Bloom's taxonomy (1956) classifies individual behaviors in three domains as cognitive, affective, and psychomotor, considering the similarities and differences. In general, properties concerning these three domains are aimed to be gained in educational institutions. Properties related to cognitive and affective domains can be named as Psychological Structure. It is obvious that the measurement process is inevitable in order to determine whether the psychological structures which are aimed to make individuals gain in school environments actualize, or individuals' levels of having these psychological properties.

As these Psychological structures are not directly measurable, developing measuring instruments which have adequate properties became a necessity. Therefore, it is clear that two basic test theories are used in order to develop measuring instruments. The first of these is known as Classical Test Theory (CTT), and the second as Latent Trait Theory (LTT) or Item Response Theory (IRT).

Since the beginning of the history of Psychological Measurement, CTT has been the most widely used theory in developing tests, analysis of test results and grading psychological scales; whereas, it can be seen that around the middle of the 20th century that Latent Trait Theory–also known as Item Response Theory–emerged and started to be used widely as an alternative to CTT due to some boundedness of CTT (Baykul, 1979; Crocker, & Algina, 1986; Gelbal, 1994; Hambleton, 1994; Kaptan, 1994; Erkuş, 2003; Reise, Ainsworth, & Haviland, 2005; Kan, 2006; Çelen, & Aybek, 2013).

The success grade in the measurement of success of individuals is obtained by combining the points that the individuals got from the items in the test according to CTT. Test and item statistics are calculated in accordance with the test points or item points of individuals. Therefore, the points that show the success levels of individuals vary in relevance with difficulty levels of the items in the test (Lord, & Novick, 1968). If the item points are not weighted considering difficulty levels of the items in the test prepared and applied according to CTT, it is accepted that items that have various difficulty levels are thought to have the same difficulty level and every single item has the same contribution level to the total points obtained according to CTT.

CTT is known as a widely used theory in order to measure some psychological structures that are admitted to be owned by individuals and to grade the measurement results obtained from measurement instruments used in order to determine the levels of these psychological structures. Furthermore, another theory, known as Item Response Theory (IRT), is accepted to become more popular and wider used than CTT, with its usage increasing day-by-day thanks to developments in technology and as a result, computer technology (Çıkrıkçı-Demirtaşlı, 1998; Reise et al, 2005; Demirtaşlı, & Arıkan, 2009; Bulut, & Kan, 2012).

It is clear that a specific theory on test development has emerged since the beginning of the 1970's. The main idea of this theory is to use fewer test items in order to measure the skills and sufficiency levels of the individual being evaluated. Through this application, skills and sufficiency levels of individuals were aimed to be determined in a more reliable way.

When Classical Test Theory and Item Response Theory are compared, the following facts may possibly be seen:

Since the item discrimination index and item difficulty index depend on the structure of the group on which the test is applied, the item difficulty and item discrimination indices of the items in the test developed according to CTT vary as the group changes. On the other hand, since IRT is independent from the group, item parameters do not change. This invariance may be seen as the superiority of IRT over CTT.

It can be said that according to CTT, providing invariance of item parameters depends on pretesting an application which is made in order to gather item parameters and some specific properties of the group on which the application is made (Hambleton, & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Hambleton, 1994; Kelecioğlu, 2001; Çelen, 2008).

While measurement errors are calculated for the whole group and standard error of the measurement is accepted as on the same level for all respondents in CTT, they are calculated individually in IRT. Therefore the individual's error level is not stable, and varies depending on the estimated skill levels of individuals. This feature may be taken as another superior fact of IRT over CTT.

Item Response Theory grounds on the information at item level more than Classical Test Theory does. Thus, which item is more distinctive on which skills level with item parameters estimated for each item, on which individuals of skill levels the measurement instrument provides more coherent measurement results and the necessary skill levels to answer the items of various difficulty levels can be estimated clearly in Item Response Theory (Nartgün, 2002 as cited in: Çelen, 2008).

While developing a test according to CTT, after pretesting the application, items of which discrimination strength is high and in order to make the item variance the highest, items of middle strength are taken into the final test. In IRT, the exact criterion is the fit of the model as well as item parameters. Furthermore, item information function, which shows the contribution of each item to the test, can be taken as a criterion in item selection (Van der Linden, & Hambleton, 1997 as cited in: Çelen, 2008).

There are several studies questioning which of these two theories mentioned above is superior to the other. When the studies made in Turkey are taken into consideration, it can be seen that those studies either investigated the exam's coherence with IRT models using a broad scale exam data or compared parameters obtained from two theories through a dataset (Can, 2003; Çalışkan, 2000; Çelen, 2008, as cited in: Kılıç, 1999; Yalçın, 1999; Çelik, 2001; Karataş, 2001; Özkurt, 2002; Yapar, 2003; Yeğin, 2003; Çelen, & Aybek, 2013).

Among the weakest points of IRT are the need for large sample groups, complexity of the theory, difficulty of interpreting results, the need for specific software, and the need for some hypotheses which are more difficult to provide than for CTT (Hambleton, & Swaminathan, 1984; Hambleton et al., 1991).

These two theories have common points: Both theories focus on the correctness and incorrectness of the responses given to test items. Normality hypothesis is the point at issue in both theories. DeMars (2010) stated that though normality hypothesis is not necessary for IRT models, the inability in normal distribution may cause problems in skills estimation. In addition to that, (Van der Linden, & Hambleton, 1997 as cited in: Çelen, 2008), stated that modern IRT is heavily affected by factor analysis, which requires normality hypothesis of IRT. When the normality of test points are provided, it is possible to transition between item difficulty and item discrimination indices which are obtained according to both theories (Crocker, & Algina, 1986, as cited in: Lord & Novick, 1968; Çelen, & Aybek, 2013). Measuring what a test is specifically developed for and that a response given to an item in test should not affect the other are the hypotheses which are supposed to be provided for both theories (Çıkrıkçı-Demirtaşlı, 1998).

Reliability in Classical Test Theory

When a measurement process is performed, errors whose sources are not certain merge in the measurement results. These kinds of errors are called random errors. There is a relationship between the concept of error and measurement instrument or reliability levels of measurement results obtained from it. As long as the level of random error which is assumed to merge in measurement results decreases, reliability of the test or reliability of the measurement results obtained from it increases. Turgut (1984) described reliability as the degree of being sterile from random errors in terms of measurement results; whereas, Baykul (1979) described it as the degree of closure of a measurement instrument to random errors and described the aspect of the degrees as precision, cohesion, and determination of the measurement instrument. In cases where the reliability levels of measurement results are high, it can be said that it relates the levels of closure to errors of the measurement instruments from which they are obtained.

In CTT, reliability coefficient is defined as the ratio of true score variance to the observed score variance, as shown in Equation 1.1 (Lord & Novick, 1968).

$$\boldsymbol{\rho_X = \rho_{XT}^2 = \frac{\rho_T^2}{\rho_X^2}} \qquad \textbf{(1.1)}$$

Reliability coefficient ρ_X which is given in Equation 1.1 and which is a parameter of the population, cannot be used in reliability calculations in normal conditions because of its present characteristic in practical applications, that while only the observed score variance (ρ_X^2) in Equation 1.1 is calculable, both true score variance (ρ_T^2) and error score variance (ρ_E^2) are incalculable. Since true score variance (ρ_T^2) and error score variance (ρ_E^2) cannot be calculated, the development and application of different reliability methods in order to calculate measurement results and accordingly the reliability coefficient of measurement instruments become inevitable.

The various reliability coefficient calculation methods developed according to Classical Test Theory are; *"equivalent forms method"*, which is based on two equivalent test forms formed by using different statements to measure the same property; *"test–retest method"*, which is developed based on a single test being applied to a group twice on different dates; *"split half method"*, which is applied to a single group and has the property of being able to be divided into two equal halves; and *"Kuder–Richardson KR-20"* and *"KR-21"* reliability calculation equations, which focus on coherence and relations among the item points obtained according to the application of a test.

Reliability in Item Response Theory

With detailed information obtained about the quality of questions used in accordance with item characteristic curve, which is considered an advantage of IRT, items can be used more efficiently in order to determine the skills level of an individual. When current measurement and evaluation approaches are considered with the characteristic feature of invariance, which removes boundedness of item parameters that depend on the group and the estimation of skills parameters that depend on items, it is possible to make a valid and reliable estimation on the skills of individuals (Crocker, & Algina, 1986; Hambleton et al., 1991; Kezer, & Koç, 2014).

In Item Response Theory, the concept of reliability in Classical Test Theory is replaced by the concept of item and test information functions.

Item information function is defined as:

$$\boldsymbol{I(\theta, U_i) = \frac{[P_i'(\theta)]2}{P_i(\theta).Q_i(\theta)}} \tag{1.2}$$

and test information function is defined as:

$$\boldsymbol{I(\theta) = \sum_{i=1}^{K} \frac{[P_i'(\theta)]2}{P_i(\theta).Q_i(\theta)}} \tag{1.3}$$

In the equations $P_i(\theta)$ (Eq. 1.2) and (Eq. 1.3) shows the chosen IRT probability function, $P_i'(\theta)$ shows its derivative, and $Q_i(\theta)$ is $Q_i(\theta) = 1 - P_i(\theta)$.

In the statement of the item information function defined as Equation 1.2, probability function $P_i(\theta)$ is in its variance and derivative. As derivative values increase and variance values decrease, information to be obtained from test items increases as well because fraction value increases. Furthermore, it is clear in Equation 1.3 that test information function is equal to the total of item information functions. This characteristic shows that item information function contributes to the test without dependence on factors besides the item. In fact, this characteristic shows IRT’s superiority over CTT.

In IRT, Standard Error of measurement is defined as dependent on information function and is showed in Equations 1.4 and 1.5 for item and test, respectively.

$$\boldsymbol{SE_i(\theta) = \frac{1}{\sqrt{I(\theta, U_{i)}}}} \tag{1.4}$$

$$\boldsymbol{SE_T(\theta) = \frac{1}{\sqrt{\sum I\,(\theta, U_{i)}}}} \tag{1.5}$$

As clear in Equations 1.4 and 1.5, standard error of measurement is inversely proportional to information functions. As the information that item and test provides increases, standard error decreases. Conversely, as the information that item and test provides decreases, standard error increases. Calculation of test information function and the fact that standard error of the test is calculable make it possible for the estimation of skill to continue until a specific error boundary while estimating the skills and sufficiency of an individual. This property is significant in order to keep the error stable in skills estimation.

Individualized Tests

Up until this point, CTT and IRT, which have different characteristics, have been briefly explained. Apart from these two theories, there is another method which can be perceived as test development and test application. This method is known as individualized (or computer-based) tests. Individualized tests are the test application methods based on application by choosing items among several questions with psychometric properties determined beforehand, according to the skill levels and properties of respondents. When individualized test is taken into consideration, contrary to paper and pencil tests, it is known as a particular test adjusted for an individual according to the skills of the individual. Individuals of different skill levels answer different items in individualized test applications, therefore the test becomes particular to the individual. In terms of this characteristic, individualized tests differ from paper and pencil tests. In paper and pencil tests, all individuals are supposed to answer the same items no matter what their level of property being measured.

In individualized tests, depending on the correctness/incorrectness of the respondent's answer, measurement process continues by giving different items which have different item difficulty levels (b_g) from an item pool. It can be said that the items determined according to this idea make it possible to produce more rational and effective skills measurement values in order to determine the skills level being measured. The next item to be answered by the respondent is determined according to the correctness/incorrectness of the item the individual responds to. If the respondent answers the item correctly, then the next item is determined as a slightly more difficult item than the (previous) one just answered. However, if the item was answered incorrectly, then the next item chosen would be easier than the (previous) one just answered.

Due to the boundedness stated about conventional tests, tailored tests or adaptive tests became a need (Lord, 1970; Weiss, 1983; Hambleton, & Swaminathan, 1985). In individualized tests, adjustment of the difficulty of test items according to an individual's skills was aimed to be measured. By means of adjusting difficulty levels according to an individual's skills, the individual can avoid getting items which are above or below their skills level (Eroğlu, 2013).

While answering individualized tests, a specific test algorithm is followed. The test algorithm; (I) start test, (II) continue answering, (III) finish answering, is given in Figure 1.

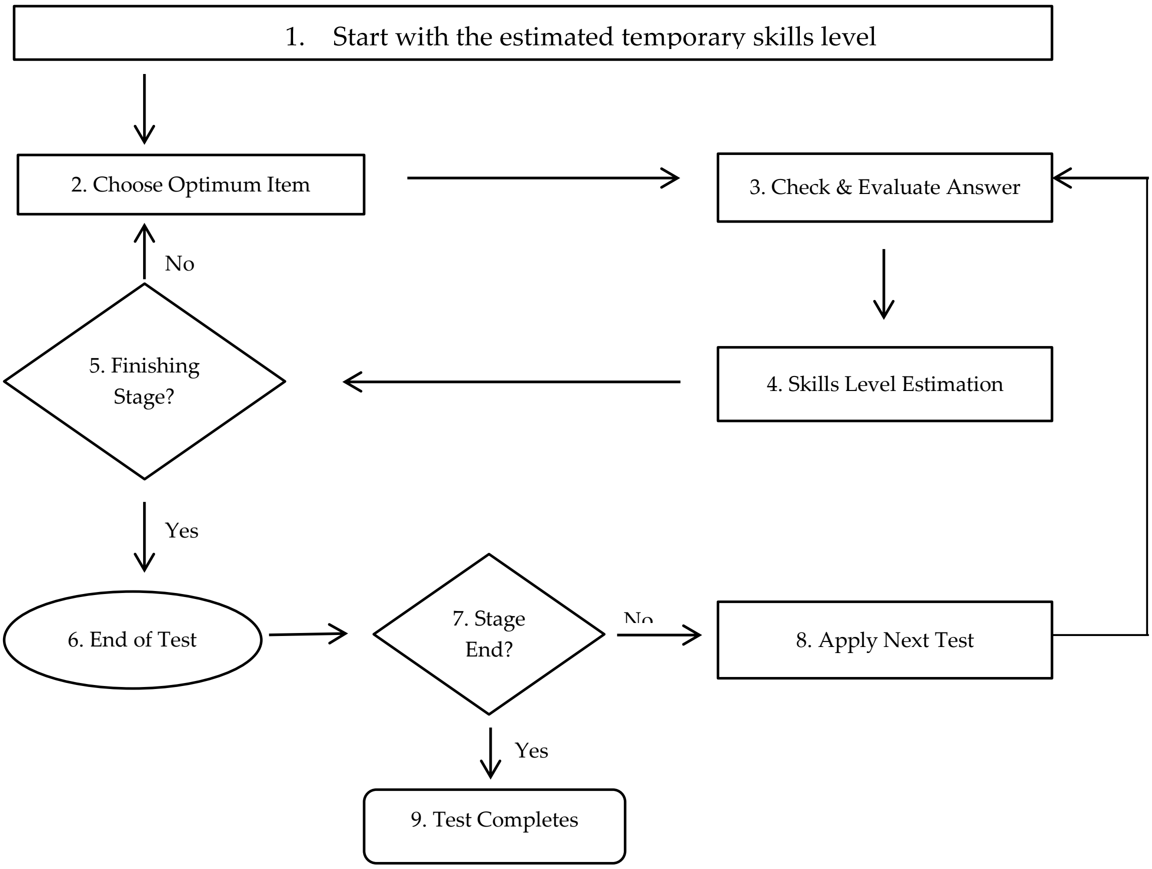

Figure 1. *Individualized Test Taking Algorithm*

The primary research question of this study is stated as:

Is there a significant difference between KR-20 reliability coefficient of the individualized test, developed using two-parameter logistic model according to Classical Test Theory, and Item Response Theory in Turkish Reading Comprehension field, and skills measurements obtained according to these two methods?

Moreover, the sub-problems to be questioned are stated as:

1. Is there a significant difference among the KR-20 reliability coefficients of the individualized test, developed using two-parameter logistic model according to Classical Test Theory, and Item Response Theory in Turkish Reading Comprehension field?
2. What level is the relation among the skills measurement obtained according to the individualized test, which is developed using two parameters logistic model according to Classical Test Theory, and Item Response Theory in Turkish reading comprehension field?

METHOD

This study is a research designed according to the basic research model. Basic researches are said to be the research models which aim to add new information to the present theoretical information of a field. This kind of research can be viewed as the studies in which the concept of research is defined in a simple way.

Since this study is in the model of basic research, the population and sampling of the research are not emphasized.

Research Data

While preparing the data collection instrument, multiple-choice test items prepared based on reading comprehension, sentence completion, synonyms, idioms and proverbs were prepared with the help of the views of two field experts. After generating 132 field-specific multiple-choice items, pretesting was applied to 132 second-grade high school students at a high school in the Denizli province of Turkey. An individualized test pool was tried to be constituted with data obtained from 134 items applied as a pilot trial by using BILOG packaged software. Following the pilot application's analysis, a total of 34 items which did not carry the required properties were excluded from the test. A total of 61 items from the remaining 100 were selected, and arranged in 13 steps based on individualized test. Application was according to the format in which each respondent would answer the 13 items from the pool deemed most suitable to their own level, take step-by-step until the test taking finished. In other words, 13 items were used as the end-point rule for this individualized test. The answering algorithm of the test application is presented in Figure 2.

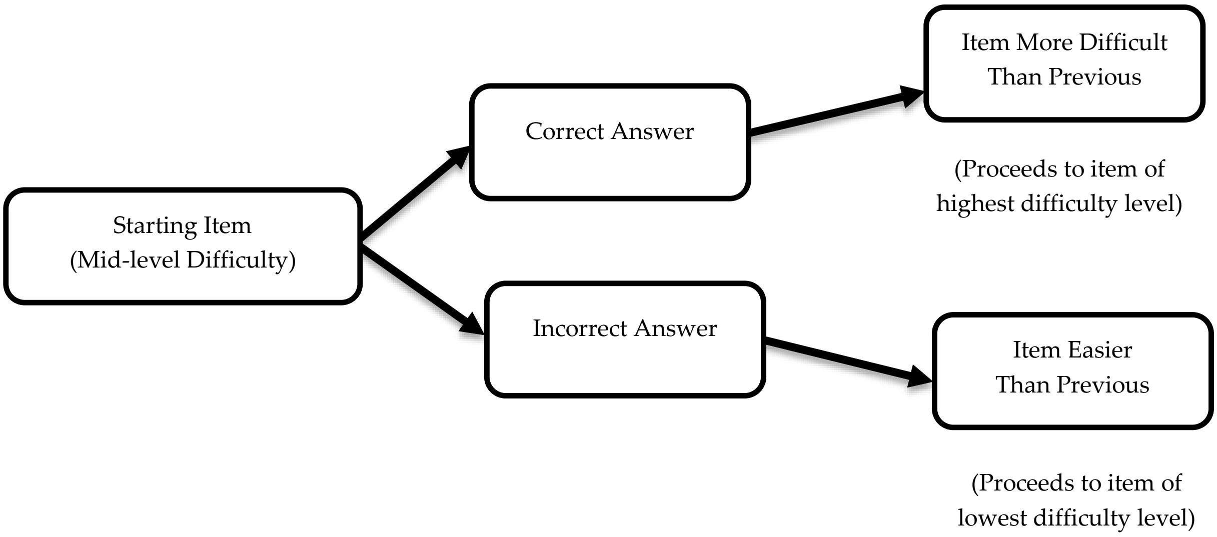
Item More Difficult Than Previous
Correct Answer
(Proceeds to item of highest difficulty level)
Starting Item (Mid-level Difficulty)
Incorrect Answer
Item Easier Than Previous
(Proceeds to item of lowest difficulty level)

Figure 2. *Answering Algorithm of Individualized Test*

It was provided that, with the help of specific software written in BASIC for the application of the individualized test, each respondent entered the program using their own password in order to answer the questions. After the respondent enters their password, the first question appears on their screen. If the respondent's answer was incorrect, another item that is easier than the first item appears on the screen which the respondent tries to answer. Or, if the respondent answered the item correctly, another item that is harder than the first item appears on the screen and the respondent tries to answer this new item.

After the respondent answers the final item in the test, the test proceeds to the finishing stage. In the test finishing stage, the answers which the respondent gave are saved to a file. After the respondent finishes the answering process, calculation of the values of skills measurement becomes easy.

FINDINGS

In order to find an answer to the first sub-problem of the research: *Is there a significant difference among the KR-20 reliability coefficients of the individualized test, developed using two-parameter logistic model according to Classical Test Theory, and Item Response Theory in Turkish Reading Comprehension field?*, the points obtained from the individualized test were subjected to linear transformation with the help of Equation 1.6:

$$X_{\theta} = 9,1\, x\theta + 100 \qquad (1.6)$$

The reason for this linear transformation of the points obtained from the test is that the signs of the individualized test points of some respondents were negative (–) and to avoid calculation problems, linear transformation process of that negative (–) sign was performed in order to calculate correlation level (Pearson product-moment correlation coefficient) between the individualized test and a paper and pencil test. Raw scores obtained from individualized test and paper and pencil test are as follows.

Table 1. *Individualized and paper and pencil test scores*

Respondents	*Individualized Test Raw Scores*	*Individualized Test Scale Scores*	*Paper & Pencil Test Score*	*Respondents*	*Individualized Test Raw Scores*	*Individualized Test Scale Scores*	*Paper & Pencil Test Score*
1	-0.19	98.29	27.48	**36**	0.48	104.34	28.77
2	-3.01	72.58	22.67	**37**	0.03	100.28	30.67
3	-0.21	98.06	31.71	**38**	0.70	106.40	29.40
4	-0.32	97.07	32.17	**39**	1.54	114.00	36.64
5	-0.36	96.76	32.32	**40**	1.22	111.12	25.59
6	1.50	113.66	24.41	**41**	-1.32	87.96	30.27
7	-0.43	103.90	28.97	**42**	-1.40	87.24	29.40
8	-0.62	94.32	33.46	**43**	-0.14	98.77	31.38
9	-2.57	76.53	26.57	**44**	0.16	101.47	30.11
10	1.35	112.24	36.49	**45**	-0.83	92.41	34.35
11	1.61	114.64	23.45	**46**	-1.93	82.44	29.42
12	2.02	118.42	42.24	**47**	-2.17	80.22	26.68
13	-1.24	88.69	36.10	**48**	0.17	101.58	30.06
14	-2.15	80.47	26.48	**49**	0.13	101.18	30.25
15	0.05	100.42	32.81	**50**	1.75	115.97	34.03
16	-2.40	78.20	24.72	**51**	0.37	103.36	31.27
17	0.47	104.28	34.45	**52**	-0.10	99.13	23.92
18	-1.40	87.29	24.65	**53**	0.28	102.53	29.62
19	0.01	100.05	30.77	**54**	0.54	104.92	28.50
20	0.22	101.99	29.87	**55**	1.80	116.42	34.74
21	0.94	108.58	33.74	**56**	0.70	106.34	31.45
22	-1.30	88.17	30.93	**57**	0.15	101.37	30.16
23	-0.50	95.50	32.91	**58**	-0.81	92.65	21.87
24	0.94	108.58	31.17	**59**	1.15	110.48	30.80
25	0.25	102.23	32.34	**60**	-0.24	97.78	20.27
26	0.50	104.52	31.29	**61**	-0.22	97.99	31.74
27	-0.76	93.05	27.62	**62**	-1.95	82.27	19.23
28	-0.26	97.64	31.90	**63**	-0.37	96.66	27.12
29	0.32	102.91	37.23	**64**	-0.87	92.12	29.10
30	-0.62	94.37	33.44	**65**	0.87	107.89	38.94
31	0.08	100.69	30.48	**66**	-0.57	94.80	30.28
32	-0.38	96.53	32.42	**67**	0.53	104.82	31.42
33	0.36	103.29	29.26	**68**	2.20	120.00	31.82
34	1.00	109.10	42.21	**69**	-2.42	87.02	19.49
35	-0.17	98.44	31.53				
				$\bar{X}$	-0.10	99.08	31.30
				S_X	1.15	10.52	10.24

According to the two parameter logistics model of the items of the Individualized test, after a_j and b_j parameter values were converted into item discrimination coefficient as item difficulty index and biserial correlation coefficient by using the equations (1.7) and (1.8) for conversion,

$$\boldsymbol{r}_{jx} = \frac{\boldsymbol{a_j}}{\sqrt{1+\boldsymbol{a}_j^2}} \qquad \textbf{(1.7)}$$

and

$$\boldsymbol{P_j} = \frac{\mathbf{1}}{\sqrt{\mathbf{2}\boldsymbol{\pi}}} \int_{\frac{\boldsymbol{a_j}.\boldsymbol{b_j}}{\sqrt{1+\boldsymbol{a}_j^2}}}^{\pm\infty} \boldsymbol{e}^{\frac{1}{2}\boldsymbol{Z}^2}.\boldsymbol{dz} \qquad \textbf{(1.8)}$$

P_j item difficulty index of item j according to CTT
r_{jx} item discrimination index of item j according to CTT
a_j discrimination index parameter of item j according to IRT

item reliability coefficients according to CTT were calculated with the help of Equation 1.9 using those values. Item reliability coefficients that were obtained are presented in Table 2.

$$\boldsymbol{r_j} = \boldsymbol{r}_{jx}\sqrt{\boldsymbol{P_j}.\boldsymbol{q_j}} \qquad \textbf{(1.9)}$$

P_j item difficulty index of item j according to CTT
r_j item reliability coefficient of item j according to CTT
q_j 1 - P_j

Table 2. *CTT Item Statistics Equivalents of Parameters of Items in Individualized Test Application*

Item No	a_j	b_j	p_j	r_{jx}	*Item No*	a_j	b_j	p_j	r_{jx}
1	0.41	0.11	0.52	0.38	**32**	0.73	-0.58	0.66	0.59
2	0.47	0.13	0.52	0.43	**33**	0.82	-0.69	0.53	0.63
3	0.65	0.12	0.53	0.55	**34**	0.69	-1.27	0.61	0.57
4	0.52	-0.20	0.54	0.46	**35**	0.78	0.56	0.63	0.62
5	0.72	-0.32	0.58	0.58	**36**	0.57	0.79	0.67	0.50
6	0.49	0.12	0.52	0.44	**37**	0.94	0.22	0.57	0.69
7	0.69	0.58	0.63	0.57	**38**	0.93	0.20	0.76	0.68
8	0.82	0.31	0.58	0.63	**39**	0.86	-0.19	0.63	0.65
9	0.74	-0.13	0.53	0.60	**40**	0.72	0.18	0.65	0.58
10	1.03	0.46	0.62	0.72	**41**	1.06	0.28	0.45	0.73
11	0.93	-0.75	0.70	0.68	**42**	0.94	0.22	0.55	0.68
12	0.83	-0.27	0.57	0.64	**43**	1.15	-0.16	0.43	0.76
13	0.67	1.05	0.72	0.56	**44**	1.07	0.86	0.41	0.73
14	1.12	0.25	0.57	0.74	**45**	0.41	0.99	0.57	0.38
5	0.78	1.58	0.83	0.62	**46**	0.77	0.43	0.55	0.61
16	0.48	-0.53	0.59	0.43	**47**	0.92	-0.12	0.42	0.68
17	0.52	-0.68	0.62	0.46	**48**	1.07	-0.78	0.72	0.73
18	0.55	1.72	0.80	0.48	**49**	0.86	0.82	0.70	0.65
19	0.66	1.24	0.75	0.55	**50**	0.53	1.68	0.78	0.47
20	0.59	0.39	0.58	0.51	**51**	0.61	0.06	0.41	0.52
21	0.47	-0.86	0.64	0.43	**52**	0.47	2.10	0.81	0.42
22	0.86	0.22	0.55	0.65	**53**	1.00	0.66	0.57	0.71
23	0.49	0.95	0.63	0.44	**54**	0.79	-0.74	0.68	0.62
24	0.73	0.19	0.54	0.59	**55**	0.98	1.03	0.67	0.70
25	1.21	-0.43	0.66	0.77	**56**	0.67	-0.86	0.76	0.56
26	0.94	-1.27	0.51	0.69	**57**	1.09	0.27	0.68	0.74
27	0.51	0.19	0.54	0.45	**58**	0.94	0.12	0.53	0.69
28	0.73	0.15	0.60	0.59	**59**	0.81	0.09	0.52	0.06
29	0.59	0.83	0.81	0.51	**60**	1.06	0.11	0.53	0.08
30	0.66	0.17	0.53	0.55	**61**	0.80	-0.57	0.63	0.36
31	0.51	0.76	0.54	0.45					

Instead of a_j and b_j values of the test item which each participant respondent in the individualized test answered, standard deviation of the test in CTT was estimated using Equation 1.10 with the converted values of a_j and b_j.

$$\boldsymbol{S_x = \sum_{k=1}^{K} r_j} \tag{1.10}$$

KR-20 reliability coefficient of the individualized test was calculated for each student by using standard deviation values and converted item statistics of the items applied to each student in the individualized test. Later, KR-20 reliability coefficient of the individualized test was calculated from the arithmetic mean of the reliability coefficients. According to CTT, KR-20 reliability coefficient was calculated as 0.75 and reliability coefficient of individualized test was calculated as 0.67.

Reliability coefficients calculated according to both methods were converted to Fisher's Z statistics with the help of Equations 1.11 and 1.12 and compared on a 0.05 level of significance. The result showed that there was no significant difference.

$$\boldsymbol{Z' = \frac{1}{2}\, ln\left(\frac{1+r}{1-r}\right)} \tag{1.11}$$

$$t = \frac{Z'_1 - Z'_2}{\sqrt{\frac{2}{N-3}}} \qquad (1.12)$$

t statistics
r correlation
Z' standard normal value of r
N number of elements in the sample

According to these results, the fact that the reliability coefficient obtained according to CTT and the reliability coefficient of the individualized test did not show a significant difference shows that both of the reliability levels are at the same level. Therefore, it does not seem possible to claim that the individualized tests meaningfully contribute to the reliability levels of items in measurement instruments or the measurement instruments of CTT. While each student answered 13 items through the individualized test algorithm, they answered 47 items in Paper and Pencil test through CTT. According to individualized testing, if the number of the questions in the item pool had been increased, the number of items which students would answer would also have increased too. In this case, the value of the reliability coefficient obtained according to individualized tests would have been greater than the reliability coefficient obtained according to CTT and it would have been possible to see a significance. Therefore, the reliability coefficient obtained by a lesser number of items in individualized tests would have been greater than paper and pencil tests formed with more items, and would have shown that individualized tests were more reliable and superior when compared to paper and pencil tests (CTT).

The second sub-problem which this research tried to answer was defined as "*What level is the relation among the skills measurement obtained according to the individualized test, which is developed using two parameters logistic model according to Classical Test Theory, and Item Response Theory in Turkish reading comprehension field?*" In order to answer address this, correlation coefficient was calculated by using Pearson product moment correlation coefficient between the students' raw scores of paper and pencil tests prepared according to CTT, and the individualized test. Correlation coefficient was calculated as 0.36 by using the raw scores of both approaches.

The reason for this low correlation coefficient obtained from the individualized test and the paper and pencil test prepared according to CTT may seem to be the biggest boundedness of this research. It should not be ignored that the lack of the desired number of items in the item pool from which the test was generated may be the cause of the low correlation coefficient obtained from the tests, which were prepared according to both methods. The fact that there were insufficient items in the pool may be related to the researcher's available resources.

CONCLUSION, DISCUSSION AND SUGGESTIONS

In this research, focusing on reading comprehension in the Turkish language, reliability levels of a paper and pencil test developed according to Classical Test Theory, and an individualized test developed according to Item Response Theory (Two Parameter Logistic Model), were compared. Because of some boundedness of the Classical Test Theory, Item Response Theory (IRT) or the theory known as Latent Trait Theory emerged towards the middle of the 20th century (Crocker & Algina, 1986). Besides the wide usage of CTT in grading the scales used in order to measure targeted properties specific to individuals, Item Response Theory (IRT) is now also increasingly used (Hambleton, 1994; Reise et al., 2005).

While the measurements obtained from CTT depend on the group, IRT is claimed to make sample-free and invariant parameter estimations. Invariance can be stated as both the invariance of the skills' parameters estimated according to the responses to different items which are prepared in order to measure the same property, and the invariance of the item parameters obtained by the application of the same test to different individuals. This fact enables a test to be used several times, so long as the property of the item does not change once it is scaled according to IRT. However, providing this invariance depends on the conditions of the trial application and the group on which this application is made (Hambleton, 1990; Hambleton, & Swaminathan, 1985; Hambleton et al., 1991; Kelecioğlu, 2001); whereas, in CTT the scores that individuals obtain vary according to the difficulty level of the test (Lord, & Novic, 1968 as cited in Çelen, 2008).

While in Classical Test Theory, standard error of measurement is calculated for the whole group, it is individually calculable in Item Response Theory; and while in Classical Test Theory, measurement errors are calculated for the whole group, it can be individually estimated in Item Response Theory. Furthermore, in Classical Test Theory, reliability coefficient is calculated as a single value for the score range of the respondent group; whereas, in Item Response Theory, it can be calculated as reliability of the item and test information functions for each item and skill level. While the single coefficient obtained in Classical Test Theory means that reliability does not change for different skill levels, it is clear that when reliability coefficient is calculated in repeated measurements, they are higher for individuals who have high levels of properties being measured. This fact shows that the measurement instrument will not have the same level of reliability for individuals of different skill levels (Nartgun, 2002 as cited in: Çelen, 2008).

Up until this research was performed, in the field of reading comprehension in the Turkish language, there had not been any studies which aimed to compare the reliability levels of the measurements of skills levels in individualized tests, which were developed by using two parameter logistic model according to Item Response Theory, and paper and pencil tests developed according to the Classical Test Theory.

The current research examined the reliability levels of paper and pencil test developed according to Classical Test Theory and individualized test developed according to Item Response Theory (Two Parameter Logistic Model), in terms of reading comprehension in Turkish. The KR-20 reliability level of the test developed according to Classical Test Theory was calculated as 0.75, and the reliability of the test developed according to the Two Parameter Logistic Model was calculated as 0.67. The reliability calculated according to both methods was converted to Fisher's Z statistics and tested at the significance level of 0.05 and no significant difference was observed.

On the other hand, correlation coefficient was calculated using Pearson product moment correlation coefficient in order to determine the level of correlation among the calculated skills measurements of individuals in a group in which the test was developed based on both methods. After the skills measurements were calculated, which the respondents obtained from the individualized test converted in order to calculate correlation coefficient, the correlation coefficient was calculated as 0.36, which signifies a low level of correlation. This situation may be the result of the low number of items in the individualized test, which actually requires a much wider item pool.

From this point of view, researchers who are willing to study this subject should ensure that the number of items in the item pool of individualized test are deemed sufficient. In addition, while this current research compared the KR-20 reliability levels of a paper and pencil test developed according to Classical Test Theory and an individualized test developed according to Item Response Theory (Two Parameter Logistic Model), studies which compare reliability levels of achievement tests should also be developed in terms of different disciplines by using the three parameter logistic method.

REFERENCES

Baykul, Y. (1979). *Örtük özelikler ve klasik test kuramları üzerine bir araştırma* (Unpublished Doctoral dissertation). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Bloom, B. (1956). *Taxonomy of educational objectives: Handbook I, The cognitive domain*. New York: David McKay.

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research, 49,* 61-80.

Can, S. (2003). *Ortaöğretim kurumları öğrenci seçme ve yerleştirme sınavı sözel bölümünün Madde tepki kuramı modellerine göre analizi* (Unpublished Doctoral dissertation). Middle East Technical University, Ankara.

Çalışkan, M. (2000). *Madde tepki kuramının (MTK) bir, iki ve üç parametreli modellerinin Milli Eğitim Bakanlığı Eğitimi Araştırma ve Geliştirme Dairesinin (MEB-EARGED) fen bilgisi başarı testi verilerine uygunluğu* (Unpublished Doctoral dissertation). Middle East Technical University, Ankara.

Çelen, Ü. (2008). Comparison of validity and reliability of two tests developed by classical test theory and item response theory. *Elementary Education Online, 7*(3), 758-768.

Çelen, Ü., & Aybek, E. C (2013). Öğrenci başarısının öğretmen yapımı bir testle klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 4*(2), 64-75.

Çelik, D. (2001). *Madde tepki kuramının (MTK) bir-, iki-, ve üç parametreli modellerinin Milli Eğitim Bakanlığı ortaöğretim kurumları öğrenci seçme ve yerleştirme sınavı testi verilerine uygunluğu* (Unpublished Doctoral dissertation). Middle East Technical University, Ankara

Çıkrıkçı-Demirtaşlı, N. (1998). Test geliştirmede yeni yaklaşımlar: Örtük özellikler kuramı – temel özellikleri, varsayımları, model ve sınırlılıkları. *Ankara Üniversitesi Eğitim Fakültesi Dergisi, 2*(28), 161-173.

Crocker, L., & Algina, J. (1986). *Introduction classical and modern test theory*. USA: CBS College Publishing.

DeMars, C. (2010). *Item Response Theory*. Oxford: Oxford University Press.

Demirtaşlı, N., & Arıkan, S. (2009). ÖİS'de MTK uygulamaları. In N. Koç, H. D. Gülleroğlu, & D. T. Coşkuner (Eds.), *I. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi Bildiri Kitabı* (pp. 225-227). Ankara: Ankara Üniveristesi.

Erkuş, A. (2003). *Psikometri üzerine yazılar.* Ankara: Türk Psikologlar Derneği Yayınları.

Eroğlu, M. G. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması* (Unpublished Doctoral dissertation). Hacettepe Üniversitesi, Ankara.

Gelbal, S. (1994). pMadde güçlük indeksi ile Rasch modelinin b parametresi ve bunlara dayalı yetenek ölçülerine üzerine bir karşılaştırma. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi,* 10, 85-94.

Hambelton, R. K. (1994). Item response theory: a broad psychometric framework for measurement advances. *Psicothema, 6*(3), 535-556.

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and application*. Kluwer, Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (2nd Ed.). California: Sage Publications, Inc.

Kan, A. (2006). Klasik test teorisine ve örtük özellikler teorisine göre kestirilen madde parametrelerinin karşılaştırılması üzerine ampirik bir çalışma. *Mersin University Journal of the Faculty of Education, 2*(2), 227-235.

Kaptan, F. (1994). *Yetenek kestiriminde adaptive (bireyselleştirilmiş) test uygulaması ile geleneksel kâğıt-kalem testi uygulamasının karşılaştırılması* (Unpublished Doctoral dissertation). Hacettepe Üniversitesi, Ankara.

Karataş, A. G. (2001). *Madde tepki kuramı (MTK) modellerini kullanarak bir İngilizce yeterlilik sınavının ölçeklendirilmesi* (Unpublished Doctoral dissertation). Middle East Technical University, Ankara.

Kelecioğlu, H. (2001). Örtük özellikler teorisindeki b ve a parametreleri ile klasik test teorisindeki p ve r istatistikleri arasındaki ilişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 20,* 104-110.

Kezer, F., & Koç, N. (2014). Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması. *Eğitim Bilimleri Araştırmaları Dergisi (Journal of Educational Sciences Research, JESR), 4*(1), 145-174.

Kılıç, İ. (1999). *Madde tepki kuramının (MTK) bir, iki ve üç parametreli modellerinin Öğrenci Seçme ve Yerleştirme Merkezi'nin (ÖSYM) Öğrenci Seçme Sınavına (ÖSS) uygunluğu* (Unpublished Doctoral dissertation). Middle East Technical University, Ankara.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing, and guidance* (pp. 139-183). New York: Harper & Row.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley.

Özkurt, S. (2002). *Madde tepki kuramı`nın (MTK) bir-, iki-, ve üç- parametreli modellerinin bir İngilizce yeterlik başarı testi verilerine uygunluğu* (Unpublished Doctoral dissertation). Middle East Technical University, Ankara.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14*(2), 95-101.

Turgut, M. F. (1984). *Eğitimde Ölçme ve Değerlendirme Metotları*. Anakara: Saydam.

Weiss, D. J. (Ed.). (1983). *New Horizons in Testing: Latent Test Theory and Computerized Adaptive Testing.* USA: Academic Press.

Yalçın, M. (1999). *Eğitimi araştırma ve geliştirme dairesi başarı testlerinin madde-tepki kuramının bir, iki, üç parametreye uygunluğu* (Unpublished Doctoral dissertation). Middle East Technical University, Ankara.

Yapar, T. (2003). *İki parametreli tepki kuramı (MTK) modelinin yetenek kestirimleriyle Başkent Üniversitesi İngilizce yeterlik sınavının yordama geçerliğini inceleme çalışması* (Unpublished Doctoral dissertation). Middle East Technical University, Ankara.

Yeğin, O. P. (2003). *Başkent Üniversitesi İngilizce yeterlik sınavının (Büiys) madde Madde Tepki Kuramı`nın (MTK) üç parametreli modelinin kullanımıyla elde edilen yetenek kestirimlerinin yordama geçerliği* (Unpublished Doctoral dissertation). Middle East Technical University, Ankara.

Please cite as:
Yaşar, M. (2017). A study on individualized tests. *Eğitim Bilimleri Araştırmaları Dergisi - Journal of Educational Sciences Research, 7*(2), 269-284. http://ebad-jesr.com/