Comparison of Item Difficulty Levels Obtained As Per Two Different Methods

Metin Yasar, PhD
Pamukkale University, Faculty of Education, Turkey

doi: 10.19044/ejes.v4no3a4 URL:http://dx.doi.org/10.19044/ejes.v4no3a4

Abstract

In this study, a multiple choice test which is composed of 19 articles which is prepared as per the scope of lesson of Measurement and Evaluation in Education, has been applied as interim exam to 207 teacher candidates who are getting education at the Faculty of Education. The difficulty levels of items which are in the test have been calculated as per Classic Test Theory. The difficulty levels of the same questions as being perceived by teacher candidates are calculated as per Classification judgment which is one of the Scaling methods and it was aimed to determine whether the difficulty leveles obtained as being based on both two methods differentiated or not. Again on the other hand, it was also determined whether there is a relation at a meaningful level between item difficulty levels being obtained as per both two methods and the direction and level of the relationship if it existed. While it was reached to the finding that item difficulty levels obtained as per both two methods differentiated even if a little, in order to determine if there is a relation between item difficulty levels obtained as per both two methods, correlation technique of Pearson and Spearman is used. Coefficient values relating with item difficulty levelts obtained as per both two correlation techniques came out to be the same, while the correlation coefficient ($r = 0{,}73$) between item difficulty levels obtained as per Classic test theory (CTT) and item difficulty levels perceived as per Classification judgments was found to be meaningful at a high level. This reveals that both two methods produced similar results with respect to item difficulty levels.

Keywords: ***Classic Test Theory, Scaling as per Classification Judgment, item difficulty level***

1.Introduction

It is required to conduct a measurement process aimining to determined whether the education received by the students attending the education programs have gained the foreseen features or not as relating with students being part of the system regardless of their levels. Since the features deter-

mined to be measures are abstract, it becomes inevitable to use indirect measurement.

In order to realize indirect measurement process, there is a requirement to have a measurement tool. Measurement processes relating with behavioral sciences require for a more meticulous study to be conducted when compared with physical measurements, as the features to be measures and the tools to be used for the measurement process are considered (Kan, 2008).

When academic success test is developed with the aim to determine the academic success of students at school (level of learning or the attained behaviors) and as the statistics of the test and the items in the test are figured out, Classic test theory (CTT) is being used. In the measurement of success, the score as per CTT, the total of scores obtained by student according to the correct answers he have to the questions in the test, reflects his success.

Test and item statistics are calculated as being based on these scores which are attained. Therefore, total of scores obtained by students as per CTT, show variations according to the difficulty level of items in the test.

Furthermore, the basic advantage of CTT is that it has got weak theoretical assumptions faciliating its being applied to many test situations (Hambleton and Jones, 1993; akt. Kan, 2006).

Due to the reason that it is easy to determine the parameters as the test is applied, applications which are based on CTT are much preferred (Kelecioğlu, 2001; Kan, 2006). In developing a test as depending on its CTT, item difficulty index (P_j) and item differentiation index are important. By using the statistics for these two items, it is being possible to estimate the features regarding the test. (Doğan, Tezbaşaran,2003; Kan,2006).

It is a known fact that the item difficulty index has the strength to influence the reliability level of the test. If the item difficulty index is very big (easiness of item) or very small (difficulty of item), this situation causes for item variances ($S_j^2 = pq$) to be small. The case where item variances ($S_j^2 = pq$) are small causes both for test reliability and item differentiation index (r_{jx}) to come out as small. Naturally, as reliability is the precondition for validility, this situation also has the power to influence the validity of the test. Item difficulty index (P_j) has influence on academic performances of students as their academic successes are measured. Especially decisions in forming certain strategies are taken about whether students who don't have sufficient level of academic success will answer to the test by looking at the difficulty levels of items in the test or not. According to Upshur (1971), measurement is an important part of education as it constantly provides information relating with the learning of students.

Furthermore, as per the researches previously conducted it is known that student-stucent interaction (Cardoso, Ferreira, Abrantes, Seabra, & Costa,

2011); success motivation, attitudes toward learning, impact of peers in learning, ethnic and gender (Abu Bakar, Ahmad Tarmizi, Mahyuddin, Elias, Wong & Mohd Ayub, 2010); education method, evaluation methods (Lebcir, Wells, & Bond, 2008); academic and general sense of self (Pullmann & Allik, 2008); intelligence and personality (Laidra, Pullmann, & Allik, 2007); exhaustion of students (Yang, 2004); peer success (Hanushek, Kain, Markman, & Rivkin, 2003); difficulty of test being perceived (Weber & Bizer, 2006; Hong, 1999) are influential on the learning performances of students.

The particulars being mentioned here also have an impact on academic success besides their influence on learning. Especially the difficulty levels tests or the items in the test being perceived, has an influence on test performance during measurement and therefore this impact is reflected on academic success. Because difficulty levels of items in the test as perceived by students, who are made subject to measurement process cause them to get worried during the exam. On the other hand, it is a known fact that there is a linear correlation between the worry felt during the exam and the academic success. Hong (1999) specifies in his research that the test difficulty (and the item at the same time) perceived during the test causes directly for the student to feel worried.

The difficulty levels perceived by students attending the exam as relating with the questions in the test, also causes them to feel worried about the exam besides causing them to feel worried as emotionally. Here the situation which brings a student to the level of worrying is the difficulty level forming in the mind of students cognitively. As per the research they conducted on 62 psychology undergratuate students, Weber and Bizer (2006) have stated that students' being informed about the difficulty of test before the application of test created a mixed impact on their performances such that it could improve or lower their performances.

However, in some of the researches made, it is being reported that there is no meaningful difference between the difficulty level of questions in a test and the academic performance of students (Laffitte, 1984; Monk & Stallings, 1970; Skinner, 1999) and even if it is stated that correction would be made for luck chance, still there was no meaningful differences between the academic success of students and the difficulty level of test items (Di- Battista, Gosse, Sinnige-Egger, Candale, & Sargeson, 2009). On the other hand, in a meta-analysis study comprising test item order as based on difficulty level, it is reported that students showed a better performance in tests starting with simply questions when compared with those starting with difficult questions or starting in a random order (Aamodt & McShane, 1992).

In psychometrics science area, in order to measure features such as success, attitude, intelligence, interest, motive, and motivation, it is needed to develop or scale measurement tools which are appropriate for the features to

be measured. Scaling pursues the goal to reveal the methods for the transition from ampirical relations to formel relations (Turgut and Baykul, 1992; Anıl and Güler, 2006). At the same time, Anıl and Güler (2006) consider scaling during the measurement process as an important ring between the passage from observations showing the qualitative differentiations and quantitative differentiations.

Scaling is analysed in two groups which are "approaches being based on trial responses and judgment decisions". The approach which is based on trial responses are centered on the respondents and it aims at the scaling of answers instead of items or stimulants (Torgerson, 1958; Turgut and Baykul; 1992; Kan, 2008; Bal, 2011). This approach focuses on placing the individuals at a different place on the scale as being based on the responses they give to the items (Crocker and Algina, 1986).

Scale development by using Likert method is one of the most well known examples of approaches based on trial responses (Tezbaşaran, 1996). These scales are the ones which are most frequently used in measuring certain features and especially attitudes in behavioral sciences (Turgut and Baykul, 1992).

The approach which is based on the decisions of adjudicator consists of scaling the stimulants as per the judgment of specialist or experts on a certain dimension and the degree of stimulation caused by each stimulant is determined with a specific method. (Ranking, classification, double comparison etc) (Stevens, 1946, cited by; Bal, 2011; Kan, 2008).Scaling approach with classification judgments is based ona statistical model aiming to determine the relations between interval limits and scaling values of stimulants in cases where the stimulants are classified in consecutive intervals. In order for gathering the judgments based on classification decisions, all of the stimulants in k num-ber are given and it is requested for it to be defined to which class each stimulant coincides with, among those classes that were priorly ranked and defined. Afterwards, as being based on judgments of observors, the scale values of stimulants are determined (Kan, 2008).

A study which compares the item difficulty levels perceived as being based on classification judgment, which is one of the scaling methods, and as per CTT of items instead of the ranking of test items within the test as having impact on the success of students, could not be specified by the researcher. This study is one which aims to compare the item difficulty levels perceived as being based on its CTT and the classification judgments, which is one of the scaling methods.

In line with this objective, it is searched to find answers to the following questions.

1-When the type of question booklet which the teacher candidates answered during the interim exam being aimed for success test for the

measurement and evaluation lesson in education is considered, are there any variations between item difficulty levels perceived as being obtained according to Classical Test Theory and those obtained as per classification judgments?

2-When the types of education applied to teacher candidates as being aimed for the achivement test for the measurement and evaluation lesson in education is considered, are there any variations between item difficulty levels calculated according to Classic Test Theory and those perceived as per Classification judgments?

3-Is there a meaningful relationship between item difficulty levels calculated according to Classical Test Theory and those perceived as per Classification judgments as being aimed for achivement test regarding measurement and evaluation lesson in education?

Research Model

Method

In this study both fundamental research and descriptive research model have been used. Fundamental researches are those researchers which add new informations to the existing theoretical informations (Kaya and Gelbal, 2007). In fundamental researches, generalization of findings obtained from the sample to the universe is not required. When it is viewed from this perspective, the study on hand can be considered as a fundamental study. Descriptive researches are those researches aiming to explain the relations between variables by considering the previous situations (Kaya and Gelbal, 2007).

Descriptive researches are generally based on survey methods. While survey methods are used to reveal the descriptive features in quantitative researches, by using the measurement tool suitable for the features being the subject of measurement, data collections are realized.

Study Group

The study group is composed of 207 teacher candidates getting education in 3rd class of Faculty of Education at Pamukkale University and taking the course of Measurement and Evaluation in Educa-tion within Spring Season of 2015-2016 academic period.

Measurement Tool

As data collection tool, an academic achivement test which was composed of 19 multiple choice questions as relating with the lesson of Measurement and Evaluation in Education and being developed with the aim to be applied in interim exam of students taking the course of Measurement and Evaluation in Education as being obligatory in the Faculty of Education is used.

Ranking could be provided by asking those answering the questions in items within the test, to put a number for ranking the difficulty level of item as being perceived inside a circle placed in front of the related item. They have made classification by ranking the items from the most difficulty one with number 1 to the easiest item by ranking with number 5.

Analysis of the Data

In the analysis of data obtained, scaling technique being based on its CTT and the classification judgments is used. In the analysis of data obtained from test being composed of 19 multi choice items that is developed and applied within the scope of Measurement and Evaluation course for interim exam, to which teacher candidates making up the research sample participated in during the period, item difficulty index (P_j) as per relevant CTT has been used.

The value regarding item difficulty index (P_j) as per its CTT, is expressed as the ratio of those answering correctly to any one item in the test to the number of those taking the exam and it is calculated as given below.

$$p=\frac{\sum x_d}{n} \qquad (1)$$

P_j = Item difficulty level
$\sum xd$ =the number of those answering correctly to the item
n = total number of those taking the exam

In the scaling method according to the classification judgments, it is asked to those answering the questions to rank the difficulty levels perceived by giving a number between 1 and 5. Test items are answered by giving a number from the most difficult (1) to the easiest one (5). As relating with scaling study with classification judgments, first of all frequency and stacked frequency matrices have been formed as relating with classification judgments obtained from adjudicators and then, stacked ratio matrix has been established from stacked frequency matrix. After this stage, by calculating unit normal deviations corresponding to each stacked ratio by using excel program, matrix for unit normal deviations (Z) is formed and with this matrix and D form, scaling process was carried out from full data matrix. By taking the averages of this matrix by columns, limit values for classes are estimated. Afterwards, general average of matrix is calculated and by subtracting the line averages from this average, scale values of stimulants (items) is predicted.

Findings and Results

In order to answer to the first sub-problem of the research, first of all item difficulty level per its CTT was calculated. As being based on

classification judgments method, in order to calculate the difficulty levels of 19 multi choice items as being perceived by answerers, first of all frequency (*F*) matrix has been formed with the aim to determine number of times a test item is placed in a class by students as being the adjudicators. In the following step, by adding the line elements of F matrix, stacked frequencies matrix is formed and by dividing frequency values in each cell of columns of matrix to the number of adjudicators, (*P*) matrix has been obtained. Unit normal deviations corresponding to each one element on the ratio matrix are calculated with the help of excel package program and matrix of unit normal deviations (Z) is formed. With regards to the unit normal deviations matrix, first of all line averages **(**Z_j**)** and column averages of the matrix **(**t_g**)** have been calculated. Column totals form the upper limit values of classes. By dividing the upper values of relevant class to the number of classes, general average of the matrix is obtained. In the next stage, by taking the difference of line averages of matrix **(**Z_j**)** from general averages of matrix **(**$\overline{Z}$**)** scale value for each test item is obtained. The smallest value of test items is taken as starting point (0.00) and by adding the absolute value of this smallest value to tge scale values of other items, new scale values as being the starting point (0.00) have been obtained. Statistics regarding the difficulty levels perceived relating with items calculated as being based on classification judgments for 19 multi choice items being part of measurement tool, are presented in table 1.

For the consistency of scale values for the difficulty levels perceived regarding the items calculated as being based on classification judgments, A.D value is calculated as 0,066. It is tested whether the model established as per the observation outcomes specified in accordance, comply with the empirical data or not. The differences between the theoretical data obtained from the model by going from the last to the beginning processes and the empirical data are compared (Kan, 2008). A.D. Data which are obtained as considered as the measure of consistency between theoretical data and empirical data. If the A.D. coefficient (value) which is obtained is small, it is considered as the consistency of indicator and if A.D. Coefficient is big, it is seen as the indicator of inconsistency of scale values. In this study, A.D. Coefficient for the classification judgments is significantly low, which shows that scale values of difficulty levels perceived for the items as being based on classification judgments are reliable.

When table 1 is investigated, while the most difficult item which is perceived as per classification judgments is the item with number 17, statistically item with number 17 has been the fourth most difficult item as per its CTT:

Again, as per item difficulty level calculated according to its CTT, the most difficult item was 8th item, whereas this item has been seen as the fifth most difficult item as per the difficulty level perceived according to classification judgments. On the other hand, the most simple item, meaning the item with less difficulty level for the answerers has been the 16th item. According to both of the methods, the answerers considered the item with number 2 as the second most difficult item. The question asked in the second sub-problem of research is related with whether the item difficulty levels obtained as per both methods differentiated or not when the type of education which the teacher candidates took was considered.

Table 1

Item difficulty levels calculated as per classification judgments and classical test theory for questions asked in interim exam for Measurement and Evaluation course

	Item difficulty level perceived as per classification judgments												**Item difficulty level perceived as per classical test theory**									
Interim Exam					Normal Education		Secondary Education		**Form A**		**Form B**		Interim exam		Normal Education		Secondary Education		**Form A**		**Form B**	
ITEM NO	z_j	s_j	**(Scale value) S_c**	**Ranking of difficulty perceived**	**(Scale value) S_c**	**Ranking of difficulty perceived**	**(Scale value) S_c**	**Ranking of difficulty perceived**	**(Scale value) S_c**	**Ranking of difficulty perceived**	**(Scale value) S_c**	**Ranking of difficulty perceived**	**Item difficulty index p_j**	**Calculated Item Difficulty Rank**	**Item difficulty index p_j**	**Calculated Item Difficulty Rank**	**Item difficulty index p_j**	**Calculated Item Difficulty Rank**	**Item difficulty index p_j**	**Calculated Item Difficulty Rank**	**Item difficulty index p_j**	**Calculated Item Difficulty Rank**
1	**-0,613**	**0,345**	**1,5594**	**15**	**1,9455**	**15**	**1,39**	**10**	**1,5594**	**15**	**1,7339**	**14**	**0,797**	**16**	**0,83**	**17**	**0,54**	**14**	**0,8116**	**16**	**0,7866**	**16**
2	**-1,144**	**0,875**	**2,1224**	**18**	**2,4766**	**18**	**2,03**	**18**	**2,1224**	**18**	**2,4767**	**18**	**0,957**	**18**	**0,98**	**19**	**0,66**	**18**	**0,9990**	**18**	**0,932**	**18**
3	**-0,433**	**0,164**	**1,4133**	**12**	**1,4666**	**10**	**1,44**	**11**	**1,4133**	**12**	**1,3938**	**6**	**0,662**	**9**	**0,69**	**9**	**0,45**	**10**	**0,757**	**13**	**0,573**	**8**
4	**-0,1888**	**-0,0881**	**1,1668**	**8**	**1,3348**	**8**	**1,30**	**8**	**1,1668**	**8**	**1,5011**	**10**	**0,502**	**6**	**0,49**	**6**	**0,37**	**6**	**0,476**	**6**	**0,534**	**6**
5	**-0,774**	**0,506**	**1,7555**	**17**	**2,0338**	**16**	**1,76**	**17**	**1,7555**	**17**	**2,1661**	**17**	**0,778**	**15**	**0,76**	**13**	**0,57**	**15**	**0,768**	**15**	**0,796**	**17**

6	0,164	-0,433	0,8169	4	1,0889	5	0,93	4	0,8164	4	1,2407	5	0,507	7	0,50	7	0,38	7	0,553	7	0,466	5
7	0,316	-0,585	0,6643	2	0,9833	2	0,89	3	0,664	2	1,100	3	0,237	2	0,29	2	0,12	1	0,204	1	0,272	2
8	0,050	-0,318	0,9311	5	1,079	3	0,94	5	0,931	5	1,058	2	0,203	1	0,16	1	0,15	2	0,262	3	0,146	1
9	0,013	-0,282	0,9677	6	1,301	6	1,12	6	0,967	6	1,456	9	0,386	5	0,30	3	0,33	5	0,233	2	0,544	7
10	-0,206	-0,062	1,1887	9	1,427	9	1,20	7	1,187	9	1,446	7	0,309	3	0,33	4	0,20	3	0,330	5	0,291	3
11	-0,443	-0,175	1,4243	13	1,657	13	1,48	14	1,424	13	1,648	12	0,749	13	0,70	10	0,58	17	0,767	14	0,738	13
12	-0,048	-0,220	1,0229	7	1,308	7	1,31	9	1,029	7	1,451	8	0,750	14	0,78	15	0,50	13	0,748	12	0,757	14
13	-0,390	-0,121	1,1370	11	1,539	11	1,47	12	1,370	11	1,580	11	0,802	17	0,80	16	0,58	16	0,845	17	0,767	15
14	-0,452	-0,184	1,4333	14	1,631	12	1,65	15	1,433	14	1,736	13	0,700	11	0,71	11	0,49	12	0,709	10	0,699	10
15	0,289	-0,558	0,6911	3	1,085	4	0,79	2	0,691	3	1,123	4	0,712	12	0,77	14	0,46	11	0,728	11	0,709	12
16	-1,151	-0,882	2,1311	19	2,539	19	2,13	19	2,131	19	2,498	19	0,966	19	0,96	18	0,69	19	0,991	19	0,951	19
17	0,980	-1,249	0,000	1	0,000	1	0,00	1	0,000	1	0,000	1	0,348	4	0,34	5	0,26	4	0,320	4	0,379	4
18	-0,691	-0,422	1,6711	16	2,214	17	1,66	16	1,671	16	2,103	16	0,681	10	0,74	12	0,43	9	0,660	9	0,710	11
19	-0,381	-0,113	1,3622	10	1,672	14	1,48	13	1,362	10	1,791	15	0,604	8	0,64	8	0,40	8	0,573	8	0,641	9

As per the finding regarding this sub-problem and as it is seen in table 1, item with number 17 has been the most difficult item both for teacher candidates having secondary education and for teacher candidates having normal education with regards to item difficulty level perceived as being based on classification judgments. On the other hand, as per its CTT item with number 17 has been seen as the most difficult fifth item for the teacher candidates having their normal education.

For the teacher candidates having secondary education program, the most difficult item was the fourth one. As per its CTT, for the teacher candidates having normal education, while the most difficult item was the 8th one, for teacher candidates having secondary education, the most difficult item was the second one. Furthermore, with regards to item difficulty level perceived as being based on classification judgments for item with number 8, for teacher candidates having normal education, the most difficult item was the third one, whereas for the teacher candidates having secondary education, the most difficult item was the fifth one.

The item having common features with regards to item difficulty levels based on both its CTT and classification judgments, has been considered to be more difficult by the teacher candidates having normal education than the teacher candidates having secondary education.

With the aim to find an answer to the third sub-problem of research, in the success test prepared within the scope of Measurement and Evaluation lesson in education, for finding an answer to the question of *"Is there a relation between item difficulty levels obtained from the calculation as per classical test theory and as per classification judgments, at a meaningful level?",* as part of 19 multi choice items, first of all the answers given by teacher candidates to test items are considered and as item difficulty levels are calculated as per its CTT and classification judgments, to be able to determine whether there is a meaningful relation between item difficulty levels obtained as per both of the methods, by using Pearson Moments Multiplication and Spearman's rho correlation technique, their levels of relationship were determined.

As the correlation coefficients obtained as per both correlation techniques were found out to be the same, only the coefficients of Spearman's tho correlation are given in table 2.

As table 2 containing the findings obtained as relating with third sub-problem of research is investi-gated, within the scope of lesson named as Measurement and Evaluation in education, as per 19 multi choice items asked in the interim exam and as regards to item difficulty levels perceived according to its KKT and classification judgments, correlation coefficients obtained as per Pearson and Spearman's tho correlation technique are calculated as being equal.

The smallest correlation coefficient (r =,67) was obtained between teacher candidates getting normal education as per CTTs and teacher candidates participating in interim exam. Smallest correlation coefficient (r=0,71) for item difficulty levelts perceived as per classification judgments, was found to be between item difficulty levels of teacher candidates getting normal education and teacher candidates answering form A in interim exam. If the correlation coefficient specified as per Pearson and Spearman's rho

technique between item difficulty levels perceived as being based on CTTs and classification judgments is small, this reveals that there were big changes in the ranking of item difficulty levels being calculated as per both two methods.

Table 2:

Statistics regarding Pearson and Spearman's rho correlation coefficients for difficulty levels as per classification judgments of academic success test items of measurement and evaluation lesson in education and for item difficulty levels calculated aps per classical test theory.

		Spearman's rho correlation									
		Interim exam for classiciation judgments					**Interm exam for classical test theory**				
		ASAG	**ASNÖ**	**ASİÖ**	**ASAF**	**ASBF**	**ASAG**	**ASNÖ**	**ASİÖ**	**ASAF**	**ASBF**
ASAG	**Correlation Coefficient**	**1,00**	**,73****	**,97****	**,95****	**1,00**	**,91****	**,68****	**,75****	**,77****	**,76****
	Sig. (2-tailed)		**,00**	**,00**	**,00**	**,00**	**,00**	**,00**	**,00**	**,00**	**,00**
ASNÖ	**Correlation Coefficient**		**1,00**	**,74****	**,71****	**,73****	**,73****	**,97****	**,97****	**,96****	**,98****
	Sig. (2-tailed)			**,00**	**,00**	**,00**	**,00**	**,00**	**,00**	**,00**	**,00**
ASİÖ	**Correlation Coefficient**			**1,00**	**,94****	**,97****	**,96****	**,70****	**,75****	**,75****	**,77****
	Sig. (2-tailed)				**,00**	**,00**	**,00**	**,00**	**,00**	**,00**	**,00**
ASAF	**Correlation Coefficient**				**1,00**	**,95****	**,92****	**,64****	**,75****	**,73****	**,74****
	Sig. (2-tailed)					**,00**	**,00**	**,03**	**,00**	**,00**	**,00**
ASBF	**Correlation Coefficient**					**1,00**	**,91****	**,68****	**,75****	**,77****	**,76****
	Sig. (2-tailed)						**,00**	**,01**	**,00**	**,00**	**,00**
ASAG	**Correlation Coefficient**						**1,00**	**,67****	**,73****	**,68****	**,78****
	Sig. (2-tailed)							**,01**	**,00**	**,00**	**,00**
ASNÖ	**Correlation Coefficient**							**1,00**	**,91****	**,93****	**,94****
	Sig. (2-tailed)								**,00**	**,00**	**,00**
ASİÖ	**Correlation Coefficient**								**1,00**	**,96****	**,95****
	Sig. (2-tailed)									**,00**	**,000**
ASAF	**Correlation Coefficient**									**1,00**	**,92****
	Sig. (2-tailed)										**,00**

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

ASAG: General interim exam; ASNO: Normal education interim exam,
ASIO: Secondary education interim exam ASAF: A form for interim exam,
ASBF: B form for interim exam

On the other hand, if the determined correlation coefficient is found to be high, it is seen that there were few changes in the ranking of item difficulty levels calculated as per both of the methods. When correlation coefficients obtained from both correlation techniques are reviewed, it is observed that there is a meaningful relation in positive direction between item difficulty levels

calculated as per CTT's and item difficulty lelvels perceived aps per classification judgments.

Conclusion and Discussion

In this study, a success test being composed of 19 multi choice items, being prepared as based on the scope of Measurement and Evaluation in Education which is an obligatory lesson in the Faculties of education, has been applied as interim exam to 207 teacher candidates and by considering the answers given to the questions by teacher candidates and by using the measurement results obtained, item difficulty levels perceived as being based on CTTs and classification judgments are calculated. Item difficulty levels obtained as per both two methods are compared

The aim in this study is to determine whether item difficulty levels calculated as per two methods are similar or not and if there is a relation between the difficulty levels or nor and whether the existing relationship is meaningful or not. When item difficulty levels calculated as per CTTs are reviewed, it is seen that the most difficult item is the 16th item (p=0,203) and it is seen that the most difficult item perceived as being based on classification judgments is the 17th article (Sc=0,00). When this question is considered, there is a differentiation observed between item difficulty level calculated as per two methods and the item difficulty level perceived. However scale values for 7th item in the test (P=0,237) are calculated (Sc=0,664) both with regards to item difficulty level according to its CTT and as per the difficulty level perceived according to classifi-cation judgment and it is seen as the most difficult second item according to both of the methods. With respect to 7th item, no differentiation is observed as relating with difficulty levels according to both of the methods. When 19th item in the test is considered as a whole, in order to determine whether there is a statis-tically meaningful relationship between item difficulty levels obtained as per the two methods, two different correlation techniques were used. (Pearson and Spearman's rho)

While statistics for correlation coefficients obtained as being based on both correlation techniques were found to be the same, correlation coefficients were calculated (r=0,733) as per item difficulty levels according to CTTs and item difficulty levels perceived as being based on classification judg-ments. This correlation coefficient is accepted as a high correlation coefficient. It can be stated that item difficulty indices predicted as per both methods were similar meaning that similar outcomes were produced. Although there is a meaningful relation between item difficulty levels calculated as per two methods (r=0,733) , since correlation coefficient calculated between item difficulty levels as obtained accor-ding to two methods is not 1.00, with regards to the challenges relating with item difficulty levels perceived as per

CTTs and classification judgments, it is possible to think that there is a differen-tiation in their relevant ranking.

Reason for this differentiation in the ranking of item difficulty levels is related with the distribution of test items in the test. When the questions in a test are ranked from the simple ones to the difficult ones and when they are randomly places, this situation causes for higher test scores to be achieved when compared with tests in which questions are ranked from the difficult ones to the simple ones (MacNichol, 1960). Changing the place of items in the test has an impact on the item difficulty levels and it is seen in this study that this would cause item difficulty levels to be differentiated (as can be seen in table 1).

This would at least differentiate the item difficulty perceived as it is seen in this study. On the other hand, Breener (1964) has considered item difficulty levels in a test and he stated that there was no meaningful difference between test performances with regards to ranking of items from the difficult to the simple ones, from the simple to the difficult ones or as being placed randomly.

Similarly in the study Lafittee (1984) conducted with regards to the test scores and difficulty levels perceived as relating with different distribution of items, he has stated that there was no influence at a meaningful level on test performances with regards to different ranking of test items and difficulty levels being perceived. As a conclusion, it is believed that item difficulty levels perceived as being based on relevant CTTs and classification judgments, which is one of the scaling methods, shall contribute to the literature.

Furthermore, as relating with correlation coefficient (r=0,733) between item difficulty levels calculated as per both methods, test (items) could be developed as making use of the difficulty levels of test items perceived as being based on classification judgments in cases where pre-application conditions may not be convenient for specifying the item statistics in developing a test. By using this method, more reliable item pools could be established.

Researches who would like to conduct similar studies, can realize studies comprising comparisons of item statistics being based on CTTs and ranking judgments as being one of the scaling methods.

Acknowledgement:

This study has been presented as verbal notification at “Cyprus International Conference on Educational Researh 2017” and it has been supported by Pamukkale University Scientific Research Division (BAP).

References:

Albayrak Sarı, A., & Gelbal, S. (2015) İkili karşılaştırmalar yargılarına ve sıralama yargılarına dayalı ölçekleme yaklaşımlarının karşılaştırılması. ***Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi***, Cilt 6, Sayı 1, Yaz 2015, 126-141.

Aamodt,M. G., & McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. ***Public Personnel Management, 21,*** **151–160.**

Bakar, K. A., Tarmizi, R. A., Mahyuddin, R., Elias, H., Luan, W. S., & Ayub, A. F. M. (2010). Relationships between university students' achievement motivation, attitude and academic performance in Malaysia. ***Procedia Social and Behavioral Sciences, 2*****(2), 4906–4910.**

Bal, Ö. (2011) Seviye Belirleme Sınavı (SBS) Başarısında etkili olduğu düşünülen faktörlerin sıralama yargıları kanunuyla ölçeklenmesi. ***Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi,*** Kış 2011, 2(2), 200-209

Baykul, Y. (2000). ***Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması***. Ankara: ÖSYM Yayınları

Cardoso, A. P., Ferreira, M., Abrantes, J. L., Seabra, C., & Costa, C. (2011). Personal and pedagogical interaction factors as determinants of academic achievement***. Procedia Social and Behavioral Sciences, 29, 1596–1605.***

Chang, S. L. (2015) Students' perceived test difficulty, perceived performance and actual performance of oral tests. ***Social Sciences & Humanities.*** *23 (4): 1225 – 1242 (2015)*

DiBattista, D., Gosse, L., Sinnige-Egger, J., Candale, B., & Sargeson, K. (2009). Grading scheme, test difficulty, and the immediate feedback assessment technique. ***The Journal of Experimental Education, 77,*** **311–336.**

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement**?** ***Journal of Applied Econometrics, 18,*** **527–544.**

Hong, E. (1999). Test anxiety, perceived test difficulty, and test performance: Temporal patterns of their effects. ***Learning and Individual Differences, 11*****(4), 431–447.**

Kan, A. (2006) Klasik test teorisine ve Örtük özellikler teorisine göre kestirilen madde parametrelerinin karşılaştırılması üzerine ampirik bir çalışma. ***Mersin Üniversitesi Eğitim Fakültesi Dergisi, Cilt 2, Sayı 2, Aralık 2006, ss. 227-235. Mersin University Journal of the Faculty of Education, Vol. 2, Issue 2, December 2006, pp. 227-235***

Kan, A. (2008) Yargıcı kararlarına dayalı ölçekleme yöntemlerinin karşılaştırılması üzerine ampirik bir çalışma. ***Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 35: 186-194 [2008]***

Kaya, Z. & Gelbal, S. (2007) ***Eğitim bilimlerinde Yöntem*** (Ed., Demirel, Ö. & Kaya, Z. : Eğitim Bilimine Giriş). 2. Baskı, PEGEM-A Yayıncılık. Ankara

Laidra, K., Pullmann, H., & Allik, J. (2007). Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. ***Personality and Individual Differences, 42,* 441–451.**
Laffitte, R. G. Jr. (1984). Effect of item order on achievement test scores and students' perception of test difficulty**. *Teaching of Psychology, 11,* 212–213.**
Lebcir, R. M., Wells, H., & Bond, A. (2008). Factors affecting academic performance of international students in project management courses: A case study from a British Post 92 University. ***International Journal of Project Management, 26,* 268–274.**
MacNichol, K., (1960). Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished manuscript, ***Educational Testing Service***, Princeton, NJ.
Monk, J. J., & Stallings, W. M. (1970). Effects of item order on test scores. ***Journal of Educational Research, 63,* 463–465.**
Öztürk, N., Özdemir, S. & Gelbal, S. (2011). İki farklı ölçekleme yaklaşımından elde edilen ölçek değerleri tutarlılığının incelenmesi. ***20. Ulusal Eğitim Bilimleri Kurultayı. Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi.*** 8-10 Eylül 2011. Burdur.
Pullmann, H., & Allik, J. (2008). Relations of academic and general self-esteem to school achievement. ***Personality and Individual Differences, 45,* 559–564.**
Skinner, N. F. (1999). When the going gets tough, the tough get going: Effects of order of item difficulty on multiple choice test performance. ***The North American Journal of Psychology, 1,* 79–82.**
Tezbaşaran (1996). ***Likert Tipi Ölçek Geliştirme Klavuzu***. Ankara: Türk Psikologlar Derneği Yayınları.
Turgut, M. F., ve Baykul, Y. (1992). **Ölçekleme Teknikleri**. Ankara: ÖSYM Yayınları
Upshur, J. A. (1971). Objective evaluation of oral proficiency in the ESOL classroom. ***TESOL Quarterly, 5*(1), 47–59.**
Weber, C. J., & Bizer, G. Y. (2006). The effects of immediate forewarning of test difficulty on test performance. ***The Journal of General Psychology, 133*(3), 277–285.**
Yang, H. J. (2004). Factors affecting student burnout and academic achievement in multiple enrolment programs in Taiwan's Technical-Vocational Colleges. ***International Journal of Educational Development, 24,* 283–301.**