IJATE

Volume 5 Issue 1 2018

International Journal of Assessment Tools in Education

International Journal of
Assessment Tools in Education

http://ijate.net/

e-ISSN: 2148-7456

e-ISSN 2148-7456

http://www.ijate.net/index.php/ijate/index

Volume 5 Issue 1 2018

Dr. Izzet KARA
Editor in Chief
International Journal of Assessment Tools in Education (IJATE)

Pamukkale University,
Education Faculty,
Department of Mathematic and Science Education,
20070, Denizli, Turkey

Phone : +90 258 296 1036
Fax : +90 258 296 1200
E-mail : ijate.editor@gmail.com

Publisher : İzzet KARA
Frequency : 2 Issues per year
Online ISSN: 2148-7456
Website : http://www.ijate.net/index.php/ijate
Design & Graphic: IJATE

Support Contact
Prof. Dr. İzzet KARA
Journal Manager & Founding Editor
Phone : +90 258 296 1036
Fax : +90 258 296 1200
E-mail : ikara@pau.edu.tr

International Journal of Assessment Tools in Education (IJATE)

IJATE will be published biannual (one volume per year, two issues per year -January and July). IJATE welcomes the submission of manuscripts that meets the general criteria of significance and scientific excellence.

There is no submission or publication process charges for articles in IJATE.

IJATE is indexed in:

- Emerging Sources Citation Index (ESCI) (Web of Science Core Collection)
- TR Index (ULAKBIM),
- DOAJ,
- Google Scholar,
- Index Copernicus International
- Türk Egitim İndeksi,
- Open Access Journals,
- Akademik Dizin,
- Academic Keys,
- CiteFactor (ASJ),
- SIS (Scientific Index Service) Database,
- SCIPIO (Scientific Publishing & Information Online),
- MIAR 2015 (Information Matrix for Analysis of the Journals),
- I2OR Indexing Services,
- JournalTOCs
- Sosyal Bilimler Atıf Dizini (SOBIAD)
- International Innovative Journal Impact Factor (IIJIF)

e-ISSN: 2148-7456

e-ISSN: 2148-7456

Table of Contents

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Investigating the Impact of Missing Data Handling Methods on the Detection of Differential Item Functioning

Hüseyin Selvi, Devrim Özdemir Alıcı

To cite this article: Selvi, H., Özdemir Alıcı, D. (2018). Investigating the impact of missing data handling methods on the detection of differential item functioning. *International Journal of Assessment Tools in Education*, 5(1), 1-14. DOI: 10.21449/ijate.330885

To link to this article: http://ijate.net/index.php/ijate/issue/view/13
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 1-14

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Investigating the Impact of Missing Data Handling Methods on the Detection of Differential Item Functioning

Hüseyin Selvi[*1], **Devrim Özdemir Alıcı**[*2]

[1]Mersin University, Medical Faculty, Medical Education Department, Turkey
[2]Mersin University, Faculty of Education, Department of Measurement and Evaluation in Education, Turkey

Abstract: In this study, it is aimed to investigate the impact of different missing data handling methods on the detection of Differential Item Functioning methods (Mantel Haenszel and Standardization methods based on Classical Test Theory and Likelihood Ratio Test method based on Item Response Theory). In this regard, on the data acquired from 1046 candidates who entered to Foreign National Student Exam (FNSE) held in year 2016 by Mersin University (MEU) and answered Basic Skills subtest, using different missing data handling methods, differential item functioning analyses with Mantel Haenszel, Standardization and Likelihood Ratio Test methods are performed. Basic Skills test consists of 80 multiple choice items. The items are all binary scored (1-0) items. Among the participants 523 are female and 523 are male. The findings showed that the number of items flagged as DIF has changed with the used missing data handling methods. The DIF detection methods based on Classical Test Theory are more consistent within themselves compared to DIF detection method based on Item Response Theory, whereas the used missing data handling methods differentiate the DIF detected items and this difference reaches a significant level for Mantel Haenszel method

ARTICLE HISTORY

Received: 31 March 2017
Revised: 30 June 2017
Accepted: 23 July 2017

KEYWORDS

Differential Item Functioning DIF; Test and Item Bias, Missing Values; Imputation of Missing Data; Mantel Haenszel; Likelihood Ratio Test

1. INTRODUCTION

Even if the reliability of the measurements acquired with a measurement tool is investigated with different method, in some cases where the desired quality (latent trait) to be measured is mixed with other qualities, the individuals in different subgroups can be affected systematically from this situation. In the current literature it is named as “bias” and causes negative effect on validity due to the definition, and it decreases somehow the reliability.

This study was presented as an oral presentation at 2016 international 5. Measurement and Evaluation Conference at Antalya.

*Corresponding Author E-mail: hsyn_selvi@yahoo.com.tr

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.330885**

Bias that occurs as a systematic variation source and affects the validity is defined as "the difference between the probabilities of correct answer of the individual within different subgroups with the same ability level (Angoff, 1993).

From this definition, in the studies regarding the determination of the bias initially, it is understood that it is necessary to match the individuals in different subgroups regarding the ability levels and to examine statistically the item parameters of these individuals. This situation is defined as the examination of whether there is Differential Item Function (DIF) in the items or not.

It is required that the items with detected DIF should be checked by the experts and whether the DIF is due to another source rather than the desired measured quality shall be investigated. In cases that the DIF is detected to be caused by another source than the desired measured quality, it can be convinced of that the related item(s) is/are biased (Camilli & Shepard, 1994; Zumbo, 1999).

In order to provide validity of the items detected biased, it can be said that it is proper for them to be revised in possible cases, and in impossible cases to be removed completely from the test. In fact, in the literature it is described that one of the important threats that affect the objectivity and validity of the measurement tools is the bias (Kristanjansonn, Aylesworth, McDowell & Zumbo, 2005).

Bias, besides decreases the validity, presents a preventable structure as a systematic variation source. Thus, scientists have developed significantly extensive methods regarding the detection of DIF. As examples of some frequently used ones of these methods Standardization (SPD-X), Mantel-Haenszel (M-H), Logistic Regression (LR) and Likelihood Ratio Test (LRT) methods can be given (Angoff, 1993; Camilli&Shepard, 1994; Osterlind, 1983).

However, it is possible to say that nearly all of these frequently used methods and other methods have different weaknesses and strengths and many methods are developed to fix weakness of each. Hence, in DIF detection there are many different distresses like in methods acting over item difficulty (pj) index, 'pj' values are affected from the average group differences and item discrimination index (rjx). In methods based on variance analysis, variance to be affected from pj and rjx values, in methods based on correlation, 'rjx' is able to be able to process in similar ways for the groups and even if the 'pj' differs, in this case to increase correlation coefficient, the correct response likelihood of the item to operate in favor of the same group for all ability levels and non-uniform DIF situation to arise etc. (Selvi, 2013).

In addition to these in the literature, studies are showing the different DIF detection methods also being affected from many variables like number-ratio of items with DIF, test length, DIF level, sample size, DIF structure in items, and item scoring method etc. (Camili & Shepard, 1994; Gelin & Zumbo, 2003; Gierl, Jodoin & Ackerman, 2000; Narayanan & Swaminathan, 1994; Osterlind, 1983; Padilla, Hidalgo, Benitez & Gomez-Benito, 2012; Selvi, 2013).

Another variable that can change the findings acquired by the DIF detection methods is thought to be the problem of missing data. Hence, many statistical methods used today based on complete data matrix and missing data rate being increased may cause these methods to give erroneous results (Bernhard, Celia & Caotes, 1998; Molenberghs & Kenward, 2007; Woodward, Smith & Tunsatall-Pedoe, 1991).

Similarly, in the literature, including M-H, LR, SIBTEST, it is said that many DIF detection methods are not capable of handling missing data (Banks, 2015). Missing data can be formed in cases like, for a performance test not reaching the item due to time limitations, accidentally

omitting the item or leaving it empty due to not knowing the right answer (Banks, 2015); for a scale, accidentally omitting the related item or refusal to answer due to personal reasons. In other words, and in the most general sense, the missing data can be considered as an information loss (Alpar, 2011).

Missing data may lead to problems like decrease of the power of the used statistical analyses, faulty estimate of standard error, increase in Type I error rate, not being able to estimate in quality the closed properties based on observation (Hohensinn & Kubinger, 2011; Molenberghs & Kenward, 2007). Thus, many studies have been done in line with the resolution of the missing data problem in time and many different methods have been developed.

Regarding the proper method to be chosen, primarily the pattern and the mechanism of the missing data should be understood. For this aim the issues like whether the missing data is distributed over the observations randomly, whether they have a specific pattern, how much missing data there is (how frequently it occurs) etc. are investigated. In other words, it is researched whether there is a case leading to missing data process in the data or not is researched (Alpar, 2011). In the literature regarding this process, it is mentioned that researchers acting carefully in data collecting presents an opportunity in observing the reasons and increasing the quality of the possible missing data (Pigott, 2001).

On the other hand, the researchers in general act in tendency to prove the assumption that the missing data does not make a significant difference on the study findings and can perform listwise deletion of the missing data with the assumption that it is missing at random (MAR) without investigating whether it is negligible or not (Alison, 2002; Groves, 2006).

In ignoring the missing data problem (un)consciously, it is thought that conditions like the researcher not having sufficient knowledge on the field of missing data problem, in scoring of the measuring tools where the maximum performance are measured (especially in optic reader usage) 1 point to correct answered items and 0 points to be assigned to the incorrect, left empty or different marking done items thus the missing data being removed by zero imputation method somehow without examination, in some statistical software the missing data to be removed by a default method automatically etc. are in play. This condition is specially emphasized in a study done by Demir & Parlak (2012). In the related study 405 researches conducted in Turkey universe and containing statistical analysis process are examined and in 40% of these studies, despite containing different analysis methods like standard error, mean, variance, covariance, correlation, t and F statistic, reliability and validity coefficients, factor analysis, regression analysis, structural equation modelling analyses, it is indicated that there was no explanation/proof seen regarding whether the data set on which the analyses are conducted had missing data or not. Listwise deletion and zero imputation make the resolution of the problem fairly ease in cases that the missing data is really formed as missing at random. However, any method to be used before the quality of the missing data is understood also consists of the possibility that the study findings are faulty.

Rubin (1976) defined three possible conditions regarding the understanding of the quality of the missing data (Missing Data Mechanism). These define cases in which the missing data is formed as missing completely random, MCAR, missing at random, MAR, and missing at non-random, MNAR. MCAR explains the situations that the probability of a value regarding x variable to be a missing data is not related to x variable itself or any value regarding another variable in the data set (Alison, 2002). In other words, MCAR explains the cases where there are no justified explanations is made regarding the formation of the missing data and the formation of the missing

data is referenced to randomization (Peng & Zhu, 2008). When the condition is looked at from DIF angle, Banks (2015) says that the MCAR missing data formation is realized in general when the related item is left empty both by the focus and the reference groups accidently.

MAR expresses the cases where the probability of a value regarding x variable to be a missing data is not related to x variable itself when the other variables in the data set are fixed (Alison, 2002). In other words, MAR is the cases in which the probability of missing data formation in the certain item is related to the observed data systematically. In the perspective of DIF definition, this situation is explained as for a test includes 30 items, the probability of the DIF analyzed items without any response (empty items) is dependent on which group that the individuals are in (focus, reference) or their performances in 2nd - 29th items (Peng & Zhu, 2008).

MNAR is the cases where the probability of a value regarding x variable to be a missing data is related to x variable itself. In other words, MNAR explains the cases where the probability of individuals to leave the item empty depends on the performances of individuals on the related item, item being left empty as it is faulty etc. (Peng & Zhu, 2008). Alison (2002), based on the definitions Rubin (1976) made regarding the quality of the missing data, classified the missing data simply as ignorable and nonignorable. In order for the missing data to be ignored, Alison said that it should be in MAR or MCAR and a missing data in MNAR cannot be ignored. Here, by the ignorable term means the case where extra modelling of missing data is not needed for the analyses to be made.

In the literature search regarding the missing data problem, there are many studies suggesting a resolution of this problem and many different methods have been developed. These methods in general are classified within as methods based on deletion and value assignment (Alpar, 2011; Demir, 2013; Alison, 2002; Little & Rubin, 1987). Among methods based on deletion; listwise deletion and analysis wise deletion, among methods based on value assignment (simple); zero imputation, mean substitution, assigning mean of nearby points, assigning median of nearby points and regression imputation methods are used frequently in the literature (Banks, 2015; Little & Rupin, 1987; Alison, 2002; Alpar, 2011).

In *listwise deletion method*; the observations containing one or more missing data are removed from the data.

In *analysiswise deletion method*; observation(s) or variables with missing data are removed from the analysis if only they are to be analyzed.

As seen, deletion methods appear as fairly simple approaches regarding the resolution of the missing data problem. However, removing the missing data from the observation via deletion methods can cause serious decrease in observation numbers and a sample deemed sufficient can turn into a sample with insufficient numbers. Moreover, methods based on deletion can decrease the stability of the calculated statistics, can place the validity and generalizability of the study to distress (Alpar, 2011). In addition to this, for methods based on deletion to be used, the assumption of missing data being in MCAR should be met (Alison, 2002; Alpar, 2011).

In *methods based on value assignment*, new values are assigned to the missing values based on specific assumptions and rules. In assigning these values (except zero imputation method) the other values or variables in the data set are considered.

In *zero imputation method*, omitted item is considered as ‘wrong’ or in most general state ‘zero’ points are assigned to this value. However, as this condition leads to biased parameter estimates and faulty hypothesis results, in Item Response Theory (IRT) and DIF studies it is especially not recommended (Banks & Walker, 2006; Lord, 1974).

In *mean imputation method*, empty value(s) is/are filled via taking the average of the values given by other individuals to the related item as serial mean imputation, via taking the average of the values given to other items by the individual as unit's mean imputation, via taking mean of nearby points, via taking median of nearby points etc. However, this condition too, can cause bias addition to many analysis results including variance-covariance estimates and parameter estimates (Little & Rupin, 1987). Similarly, for these assignment methods to be used the assumption of missing data being in MCAR should be met (Alpar, 2011).

The regression imputation method; is based on estimation operations realized by taking the regressed variable as the variable with missing value(s) and other variable(s) as regressing variables. However, in this method, as it starts upon relations between other variables, the already present relation in the data can be strengthened more as the result of the assignment thus lead to being biased. In addition, the value obtained as the result of estimation can exceed the score range of the missing data. In order to use the regression imputation methods, the missing data being in MCAR should again be met (Alpar, 2011).

The methods based on deletion and value assignment appear as frequently used method in resolution of the missing data problem. However, it is known that these methods also bring up many restrictions. These restrictions, whereas, drove the researchers to develop new methods.

Among the methods suggested in this regard, the multiple imputation method suggesting estimation of the missing data via using two or more methods together and Expected-Maximization method based on maximum likelihood shine out are mentioned (Alison, 2002; Alpar, 2011; Demir, 2013; Little & Rubin, 1987). The most important advantage of these methods compared to methods based on deletion and simple value assignment is that they can also be used in cases where the missing data is in MAR (Alison, 2002; Alpar, 2011).

When the studies performed in literature regarding the missing data problem and used methods are examined; it is suggested that in cases that may cause serious reduction in data set or bias listwise deletion shall not be used (Graham, 2009). As it increases Type I error rate the zero imputation method shall be avoided if possible (Banks & Walker, 2006; Banks, 2015; Robitzsch & Rupp, 2009). The method with Type I error rate that is similar to the complete data set shall be preferred (Banks & Walker, 2006; Finch, 2011) and especially in DIF studies the missing data problem shall not be ignored (Banks, 2015). Besides; it is expressed that sample size and DIF level in items being increased, the performance of analysiswise deletion methods instead of listwise deletion and zero imputation methods, increase the rate of accurately determined items with DIF. It is shown that item to grow difficult and missing data rate to be increased decreases as well (Banks & Walker, 2006; Emenogu, Falenchuck & Childs, 2010; Finch, 2011; Garrett, 2009).

On the other hand, the most efficient solution in missing data problem can be shown as with precautions like being careful yet on the data gathering stage, training individual given the task of data gathering, the missing data not to be present or be in ignorable quality and level (Alison, 2002; Little & Rubin, 1987). In this regard, there are different suggestions in literature regarding the ignorable missing data ratio. Schafer (1999) said that this rate should be below 5%, Bennett (2001) 10%, Peng, Harwell, Liou & Ehman (2006) 20% and otherwise it should be considered that the findings acquired from the study may be biased.

The missing data problem and DIF are still seen important problem and research studies on these topics are ongoing. In the literature there are many extensive studies regarding the detection of the lacking and powerful points of the missing data approaches and DIF detection methods.

However, it is observed that nearly all of these studies were performed over data sets acquired by simulation method (e.g., Banks & Walker, 2006; Banks, 2015; Emenogu, Falenchuck & Childs, 2010; Falenchuck & Herbert, 2009; Finch, 2011; Garrett, 2009; Hohensinn & Kubinger, 2011; Pigott, 2001; Robitzsch & Rupp, 2009; Rousseau, Bertrand & Boiteau, 2006; Sedivy, Zhang & Traxel, 2006). And it is observed that nearly all of these studies were performed over frequently used DIF detection methods like, Standardization, SIBTEST, Linear Logistic Regression and Likelihood Ratio Test (e.g., Banks, 2015; Finch, 2011; Robitzsch & Rupp, 2009; Wu, Lee & Zumbo, 2007). A study which includes the Classical Test Theory (CTT) and Item Response Theory (IRT) based DIF detection methods, a non-simulative data set and expected maximization and regression imputation methods at the same time is not seen.

In the literature, regarding the studies conducted on simulation technique, it is expressed that being aware of the situation that these studies cannot present enough proof that the actual results shall be found and cannot guarantee the accuracy of the results to be found and thus it is imperative to be sure exactly that all the analytic and experimental options that can be used in solving the problem would not be usable before these studies are performed and finally they should be used as last resort (Harwell, Stone, Hsu & Kirisci, 1996).

Thus in this study, the answer of the question "How are the performances of expectation maximization and regression imputation methods for handling with missing data on detecting DIF methods based on CTT and IRT is sought.

2. METHOD

In this study, over the complete data matrix obtained by using different missing data methods, the investigation of operation of DIF detection methods based on different theories in regard to gender variable is aimed for. Thus it can be said that the type of this study is basic research (Kothari, 2004; Royce, Straits & Straits, 1993; Singh, 2006).

Data acquired from 1046 candidates who attended to the Foreign National Student Exam (FNSE) conducted by Mersin University (MEU) in year 2016 and answered Basic Learning Skills subtest.

Some descriptive information related to the participants is given in Table 1.

Table 1. Descriptive values regarding the participating group

	Foreign National	Turkish National	Total
Female	448 (50.1%)	75 (49.3%)	523 (50%)
Male	446 (49.9%)	77 (50.7%)	523 (50%)
Total	894 (100%)	152 (100%)	1046 (100%)

2.1. Instrument

FNSE consists of two subtests as Basic Skills Test and a Language Test and is applied to high school graduates in Turkey and specific centers around the world every year for granting them undergraduate education in MEU. Candidates are ranked according to the scores they achieved in this exam and regarding specific quotas, are placed to programs they chose. In

development of the tests, all works are planned and realized by the Measurement and Evaluation Application and Research Centre of the university. The Basic Skills subtest was used as data collecting tool in this study is scored in binary (0-1), multiple-choice and consists of 80 items with 5 choices and the reliability (KR 20) of the acquired scores is calculated as 0.95.

2.2. Data Analysis

DIF analyses was done via M-H, Standardization and LRT methods. M-H, and Standardization methods do not contain the assumptions which of the parametric techniques should be faced. However, as LRT is one of the methods based on IRT the data should meet the unidimensionality and local independence that are basic assumptions of IRT (Embretson & Reise, 2000; Hambletton & Swaminathan, 1985). Thus, in the first stage of the data analysis whether these assumptions were checked.

In this regard, the unidimensionality that is one of the basic assumption of the IRT, is investigated utilizing the principal components analysis based on intra-item tetrachoric correlation matrix and the data is observed to be unidimensional from the acquired results regarding the local independence, in the literature it is said that this assumption is linked to the unidimensionality and a data that is seen to be unidimensional meets also the local independence (Lord, 1980: 19; Hambletton & Swaminathan, 1985: 25). Based on these it is deemed that the study data also meets the local independence.

In the second stage of the data analysis, in order the analysis based on Item Response Theory to be done, model-data fit was examined. Because the likelihood ratio test, which is one of the DIF methods used in this research based on IRT and the DIF analysis software (IRT-LR-DIF) requires the selection of the model. The -2 log likelihood value of the data obtained for the two parameter logistic model is calculated as 71207,78. As this statistic showing χ^2 distribution is very sensitive to sample size and in big sample sizes model-data fit cannot be provided for nearly all models; for evaluation of the model data fit $-2 \log likelihood / (S-1) - 2n(r-1) \leq 3.00$ condition is considered. Here '*S'* shows response pattern number, *n* number of items, *r* number of response category The possible response pattern of this study dependent on the item number and response category number is 5^{80}. Bock (1997) indicates that all values meeting the '–2 log likelihood / (S-1) – 2n(r-1) ≤ 3.00' condition are sufficient for model data fit (Gözen Çıtak, 2007). Based on these findings it can be said that the data is fit to the 2 parameter logistic model.

In the third stage of the data analysis, in order to decide the pattern of the missing data Little's MCAR test was applied and it was observed that the data was not in MCAR ($\chi 2$=22815.65, p<0.05). In the fourth stage of data analysis, the missing values that are present in the raw data set and whose ratios change in between 0.3% and 10%, due to the data not being in MCAR, are removed by Expectation Maximization and Regression Imputation and DIF analyses are made on complete data set by Mantel Haenszel, Standardization and Likelihood Ratio Test methods and items showing DIF and number of items with DIF are determined.

Whether the number of items determined with different missing data methods and different DIF detection methods show discrepancies is examined by Cohran's Q and McNemar tests. Cohran's Q test is used for testing whether the number of items with DIF determined via Mantel Haenszel, Standardization and Likelihood-Ratio Test for each missing data method, differentiate from each other or not; and McNemar test is used if there is a significant difference found by

Cohran's Q test and in order to test whether the number of DIF included items according to the used missing data method are significantly different from each other or not.

3. FINDINGS

In the scope of the study the DIF analyses performed on the complete data matrix obtained by expectation maximization and regression imputation methods and the values obtained as the result of these analyses are given in Table 2.

Table 2. Results of DIF analysis performed on complete data matrix obtained by expectation maximization and regression imputation methods.

Items	Expectation Maximization						Regression Imputation						Missing Data Ratio (%)
	Focus-Ref. Group Mean		M-H		Std.	LRT	Focus-Ref. Group Mean		M-H		Std.	LRT	
	Male	Female	$MH\chi^2$	*p*	*SPD**	G^2**	Male	Female	$MH\chi^2$	*p*	*SPD**	G^2**	
Item 1	0.9	0.87	3.67	0.05	-0.03	5.5	0.9	0.87	5.21	**0.02**	-0.03	**6.6**	1.0
Item 2	0.87	0.87	0.00	0.92	-0.01	0.1	0.86	0.87	0.09	0.75	0.00	0.10	1.3
Item 3	0.75	0.73	0.27	0.59	0.00	1.3	0.75	0.73	0.47	0.49	-0.01	1.6	1.9
Item 4	0.84	0.78	6.72	**0.01**	**-0.05**	**11.8**	0.83	0.78	7.01	**0.00**	**-0.06**	**12.2**	2.4
Item 5	0.85	0.85	0.05	0.81	0.00	1.4	0.85	0.84	0.1	0.74	0.00	1.6	1.7
Item 6	0.69	0.69	0	0.92	0	1.6	0.68	0.69	0.13	0.71	0.01	1.5	3.3
Item 7	0.44	0.41	0.39	0.53	-0.02	1.9	0.45	0.42	0.25	0.61	-0.02	2	6.8
Item 8	0.93	0.92	0	0.95	0	0.9	0.93	0.92	0.01	0.91	0	1.1	0.4
Item 9	0.62	0.63	0.02	0.87	0	0.4	0.62	0.63	0.15	0.69	0.01	0.1	7.8
Item 10	0.88	0.91	1.99	0.15	0.02	**6.9**	0.88	0.91	3.2	0.07	0.04	5.4	1.1
Item 11	0.66	0.55	19.81	**0**	**-0.12**	**22.8**	0.64	0.54	14.3	**0**	**-0.10**	**17.9**	5.3
Item 12	0.54	0.5	1.76	0.18	-0.04	**6.6**	0.54	0.51	1.12	0.28	-0.02	5.2	3.5
Item 13	0.81	0.85	1.15	0.28	0.02	3.4	0.81	0.85	1.22	0.26	0.02	4.5	1.6
Item 14	0.55	0.53	0.45	0.49	-0.02	1.8	0.56	0.54	0.46	0.49	-0.01	1.5	5.3
Item 15	0.91	0.89	0.17	0.67	-0.01	1.8	0.9	0.89	0.22	0.63	0	2	0.8
Item 16	0.82	0.86	2.98	0.08	0.02	3.7	0.82	0.86	3.22	0.07	0.03	4	1.1
Item 17	0.93	0.94	1.37	0.24	0.01	2.6	0.93	0.94	0.42	0.51	0	1.7	0.4
Item 18	0.9	0.89	0.11	0.73	0	0.8	0.9	0.89	0.07	0.78	-0.01	0.5	0.7
Item 19	0.91	0.92	0.67	0.41	0.01	2.8	0.92	0.93	0.23	0.62	0.02	2.6	0.3
Item 20	0.93	0.95	4.39	**0.03**	0.04	2.6	0.93	0.94	2.04	0.15	0.02	2.2	0.5
Item 21	0.27	0.25	0.3	0.58	-0.02	0.7	0.28	0.25	0.32	0.56	-0.02	0.2	2.2
Item 22	0.75	0.76	0.09	0.75	0.01	0	0.75	0.75	0.01	0.90	0	0	3.0
Item 23	0.66	0.63	2.1	0.14	-0.03	5.3	0.64	0.62	1.85	0.17	-0.03	3.5	6.8
Item 24	0.82	0.76	5.94	**0.01**	**-0.06**	**8.3**	0.82	0.76	5.99	**0.01**	**-0.06**	**7.1**	1.2
Item 25	0.87	0.88	0.27	0.6	0	0.3	0.87	0.88	0.56	0.45	0.02	0.8	2.2
Item 26	0.69	0.63	3.99	**0.04**	**-0.06**	**8.7**	0.69	0.62	7.85	**0**	**-0.07**	**11.3**	3.7
Item 27	0.87	0.86	0.53	0.56	0	**6.8**	0.87	0.86	0.88	0.34	-0.01	**6.4**	0.9
Item 28	0.92	0.93	0.02	0.88	0	0.5	0.92	0.93	0.14	0.70	0.01	0.5	0.9
Item 29	0.9	0.9	0	0.93	0	1.9	0.9	0.9	0.48	0.48	-0.01	2.1	1.3
Item 30	0.66	0.66	0.02	0.88	0	0.9	0.66	0.64	1.44	0.22	-0.03	1.8	5.9
Item 31	0.84	0.84	0	0.98	-0.01	0.7	0.84	0.84	0	0.93	0	1.5	1.9
Item 32	0.53	0.47	3.66	0.05	**-0.05**	**10.3**	0.53	0.46	3.91	**0.04**	**-0.05**	**10.9**	5.6
Item 33	0.69	0.72	0.73	0.39	0.03	0	0.69	0.71	0.95	0.32	0.03	0	4.3
Item 34	0.8	0.83	0.7	0.40	0.02	1	0.81	0.83	0.04	0.83	0.01	0.2	1.8
Item 35	0.64	0.65	0	0.98	0	4.8	0.64	0.64	0	0.95	0	4.6	5.0
Item 36	0.33	0.31	0.02	0.88	0	2.2	0.65	0.33	0.22	0.63	0	3.3	5.7
Item 37	0.69	0.74	3.25	0.07	**0.05**	2.9	0.69	0.73	1.82	0.17	0.03	2	5.3
Item 38	0.93	0.92	0.02	0.88	0	5.7	0.93	0.92	0.12	0.72	-0.01	**6.6**	1.0

Item 39	0.63	0.63	0.04	0.83	0	0.3	0.63	0.63	0.2	0.64	-0.01	0	3.2
Item 40	0.9	0.89	0.15	0.69	0	2.1	0.9	0.89	0.13	0.71	0	2.2	0.4
Item 41	0.7	0.7	0	0.95	0	0	0.7	0.7	0	0.95	0	0	1.2
Item 42	0.88	0.9	0.56	0.45	0.02	1.5	0.88	0.9	0.76	0.38	0.01	1.5	0.8
Item 43	0.75	0.75	0.04	0.82	0	0.7	0.75	0.75	0	0.95	0	0.6	1.8
Item 44	0.8	0.73	10.78	**0**	**-0.08**	**11.7**	0.8	0.74	9.64	**0**	**-0.07**	**8.9**	1.6
Item 45	0.47	0.42	3.73	0.05	**-0.05**	**9.9**	0.48	0.43	4.1	**0.03**	**-0.05**	**8.1**	4.4
Item 46	0.16	0.2	1.62	0.20	0.03	1.5	0.16	0.2	1.46	0.22	0.03	0.5	2.3
Item 47	0.82	0.84	0.35	0.55	0	0	0.82	0.84	0.12	0.72	0	0.1	2.4
Item 48	0.74	0.76	0.19	0.65	0.01	1.6	0.73	0.76	0.07	0.78	0	1.8	3.0
Item 49	0.71	0.78	8.08	**0**	**0.05**	**8.7**	0.71	0.79	8.61	**0**	**0.05**	**11.1**	3.1
Item 50	0.3	0.63	1.08	0.29	0.02	1.2	0.59	0.62	0.25	0.61	0.01	0.7	8.3
Item 51	0.7	0.76	2.39	0.12	0.04	4.8	0.7	0.76	4.91	**0.02**	**0.05**	5	4.7
Item 52	0.76	0.82	8.27	**0**	**0.05**	**8.4**	0.76	0.82	4.14	**0.04**	0.03	**6.2**	4.1
Item 53	0.69	0.7	0.06	0.8	-0.01	1.6	0.69	0.69	0.32	0.57	-0.01	3.6	4.8
Item 54	0.8	0.88	10.06	**0**	**0.05**	**19.1**	0.8	0.87	10.1	**0**	**0.05**	**17.4**	2.3
Item 55	0.63	0.62	1.35	0.24	-0.02	4.5	0.63	0.62	0.52	0.46	-0.01	3.7	6.6
Item 56	0.67	0.71	1.43	0.23	0.02	3.4	0.67	0.71	2.02	0.15	0.03	3.2	4.5
Item 57	0.56	0.61	1.76	0.18	0.03	2.9	0.56	0.6	1.63	0.20	0.03	2.9	6.6
Item 58	0.72	0.74	0.08	0.77	0	2.5	0.72	0.74	0.02	0.87	0	1.5	3.8
Item 59	0.55	0.61	3.78	0.05	0.04	4	0.54	0.6	4.74	**0.02**	**0.06**	5.7	6.9
Item 60	0.39	0.25	19.19	**0**	**-0.12**	**28.7**	0.39	0.26	15.8	**0**	**-0.11**	**25.9**	4.2
Item 61	0.69	0.73	2.7	0.09	0.03	2.1	0.69	0.72	1.18	0.27	0.02	2.6	5.1
Item 62	0.64	0.64	0.1	0.74	-0.02	3.7	0.63	0.63	0.07	0.78	-0.02	2.8	6.0
Item 63	0.69	0.74	3.61	0.05	0.03	2.5	0.68	0.74	4.08	**0.04**	0.03	2.9	5.2
Item 64	0.45	0.43	0.12	0.72	-0.01	4.5	0.47	0.45	0.14	0.70	-0.01	3.1	9.1
Item 65	0.76	0.76	0	0.94	0	3.2	0.76	0.76	0.01	0.89	0	1	4.8
Item 66	0.35	0.34	0.07	0.78	-0.01	0.4	0.37	0.36	0.16	0.68	-0.01	0.3	3.7
Item 67	0.15	0.17	0.04	0.83	0	3.7	0.17	0.18	0.11	0.73	0	4	10.3
Item 68	0.47	0.49	0.64	0.42	0.03	0.7	0.47	0.49	1.46	0.22	0.04	1.1	7.8
Item 69	0.49	0.5	0.07	0.78	0	5.2	0.49	0.51	0.67	0.41	0.01	1.5	9.0
Item 70	0.58	0.61	1.86	0.17	0.02	0.8	0.58	0.61	0.33	0.56	0	0.1	7.5
Item 71	0.54	0.57	0.73	0.38	0.02	1	0.54	0.57	1.17	0.27	0.03	3	8.9
Item 72	0.63	0.64	0	0.97	-0.01	**12.7**	0.64	0.64	0.18	0.66	-0.01	**15**	6.1
Item 73	0.54	0.56	0.16	0.68	0	5.2	0.54	0.56	0.19	0.65	0	3.7	8.4
Item 74	0.62	0.62	0.26	0.60	-0.01	2.6	0.63	0.63	0	0.92	0	1.9	5.5
Item 75	0.32	0.34	0.91	0.33	0.02	0.5	0.33	0.37	3.31	0.06	0.04	3.3	7.9
Item 76	0.65	0.71	2.36	0.12	0.04	2.6	0.65	0.71	3.64	0.05	**0.05**	2.6	5.8
Item 77	0.54	0.54	0.07	0.77	0	0.5	0.55	0.54	0	0.92	-0.01	0.9	7.3
Item 78	0.4	0.37	1.28	0.25	-0.02	4.6	0.42	0.39	0.35	0.54	-0.02	3.4	10.0
Item 79	0.69	0.77	6.19	**0.01**	**0.05**	**11.1**	0.68	0.75	7.27	**0**	**0.06**	**9.6**	5.8
Item 80	0.58	0.59	0.05	0.81	0.01	0	0.56	0.58	0.5	0.47	0.02	0.5	7.2

***** SPD-X values are located between '-1.00' to '1.00'. The values between -0.05 to 0.05 shows ignorable level of DIF; and values between -1 to -0.05 and 0.05 to 1 intervals shows unignorable level of DIF presence (Gonzales. Padilla. Dolores. Gomez & Benitez. 2010).

**As the G^2 values calculated with LRT test show the chi-square distribution in the freedom degree up to estimated parameter number. the critical value of the chi-square distribution here regarding the DIF detection is taken as 5.99 (p=0.05. df=2) (Dişçi. 2012).

When Table 2 is examined. in the analyses performed on the complete data matrix obtained by expectation maximization. DIF is seen in 11 items with M-H method. 13 items with Standardization method and 16 items with LRT method. Similarly, on the complete data matrix obtained by regression imputation method. DIF is seen in 16 items with M-H method. 14 items with Standardization method and 16 items with LRT method.

The results of Cohran's Q and McNemar tests performed regarding whether the items determined with different missing data methods and different DIF detection methods show difference and simple coefficient of concordance calculated related to these are shown in Table 3 and Table 4.

Table 3: The results of Cohran's Q and McNemar tests performed regarding whether the items determined with different missing data methods and different DIF detection methods show difference.

Missing Data Methods	MH. Std. and LRT		MH (em-reg.)	Std. (em-reg.)	LRT (em-reg.)
	Cohran's Q	p	McNemar (p)	McNemar (p)	McNemar (p)
Expectation Max.	4.75	0.09	0.03	1.00	0.34
Regression Imputation	0.89	0.64			

Table 4. Simple coefficient of concordance calculated related to items determined with different missing data methods and different DIF detection methods

	MH. Std. and LRT	MH-Std.	MH-Std.	Std.-LRT
Expectation Max.	0.91	0.95	0.91	0.95
Regression Imputation	0.91	0.95	0.93	0.91

When Table 3 is examined. in the analyses performed on the complete data matrix obtained by both expectation maximization and regression imputation according to the Cochran's Q test results. items determined to be with DIF are observed to be differentiated from each other significantly. McNemar's test results show that the items determined by M-H method are differentiated significantly with the used missing data method. In the other DIF detection methods examined in the scope of the study in items with DIF determined regarding the used missing data method there has no significant change occurred.

The findings acquired in the scope of the study showed that the item numbers showing the DIF are changed among the DIF detection method. the DIF detection methods that are used in the scope of the study and based on the Classical Test Theory are more fit internally compared to the DIF detection method based on IRT. the used missing data approaches differentiate the items determined to be with DIF and this difference reaches to a significant level for Mantel Haenszel method.

4. DISCUSSION, CONCLUSION AND SUGGESTIONS

The findings acquired in this study showed that the items included DIF and their numbers were changed based on DIF detection method. The findings are partially overlapping with the findings of the other studies in the literature (Abedlazeez, 2010; Doğan & Öğretmen, 2008; Finch, 2011; Hohensinn & Kubinger, 2011; Kan, Sünbül & Ömür, 2013; Pigott, 2001; Robitzsch & Rupp, 2009; Spray & Miller, 1994; Ward & Bennett, 2012). Hence, in many of these studies significant difference between the items determined with different DIF methods and their numbers are present, whereas the determined difference in this study did not reach a significant level. Among the reasons, the difference between the item difficulty values obtained from the focus and reference

groups to be very close to zero, the related items and the test to be possibly qualified as ‘easy’ by the item difficulty value averages can be shown.

On the other hand, even if there was no significant difference between the results of DIF methods used for the missing data methods, the methods based on CTT are observed to have more concordance within compared to the methods provided by the methods based on IRT. The main reason of this can be shown as the M-H and Standardization methods to be calculated over contingency table and based on the same theory. These findings are overlapping with the findings of Selvi (2013).

In addition to these it is seen from the acquired findings that the used missing data approaches differentiate the items determined to be with DIF and this difference reaches to a significant level for Mantel Haenszel method. The findings acquired are overlapping with the findings of Robitzsch and Rupp (2009). In short, based on the findings obtained in the scope of this study and related literature, the conclusion can be reached that the used missing data approach, being also dependent on the DIF detection method, differentiate/can differentiate the items determined to be with DIF.

This result shows the possibility of the findings to be erroneous of the studies in which the missing data pattern and mechanism are ignored consciously/unconsciously or an inappropriate missing data approach is chosen and this reduces the importance of the missing data problem to an extent. The findings obtained in the scope of this study are limited with the expectation maximization and regression imputation methods among missing value assignment methods; and Mantel Haenzsel, Standardization and Likelihood Ratio Test methods among the DIF detection methods. Thus it can be suggested that similar studies, considering also the variables like scoring condition, sample size, different psychometric properties of items etc., shall be repeated with different missing data assignment method. Different DIF detection methods and the operation of different missing data methods on DIF shall be examined in order to contribute in solution of the missing data problem**.**

5. REFERENCES

Abedlazeez, N. (2010). Exploring DIF: Comparison of CTT and IRT methods. *International Journal of Sustainable Development*, 7(1), 11-46.

Allison, P. D. (2002). *Missing data.* California: Sage Publication Inc.

Alpar, R. (2011). *Uygulamalı çok değişkenli istatistiksel yöntemler.* Ankara: Detay Yayıncılık.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In Holland & Wainer (Ed.), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates Publishers.

Banks, K., & Walker, C. (2006). *Performance of SIBTEST when focal group examinees have missing data.* San Francisco: National Council of Measurement in Education.

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research & Evaluation*. 20(12).

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, *25*, 464–469.

Bernhard, J., Celia, D.F., &Coates, A.S. (1998). Missing quality of life data in cancer clinical trials: Serious problems and challenges. *Statistics in Medicine*, 17, 517-532.

Camili, G., & Shepard, L.A. (1994). *Methods for identifying biased test items.* London*:* Sage Publication.

Demir, E., & Parlak, B. (2012). Türkiye'de eğitim araştırmalarında kayıp veri sorunu. *Journal of Measurement and Evaluation in Education and Psychology 3*(1), 230-241.

Demir, E. (2013). Kayıp verilerin varlığında çoktan seçmeli testlerde madde ve test parametrelerinin kestirilmesi: SBS örneği [Item and test parameters estimations for multiple choice tests in the presence of missing data: The case of SBS]. *Journal of Educational Sciences Research, 3*(2), 47–68.

Dişçi, R. (2012). *Temel ve klinik biyoistatistik.* İstanbul: Tıp Kitapevi.

Doğan, N., & Öğretmen, T. (2008). Değişen Madde Fonksiyonunu belirlemede Mantel–Haenszel, Ki-Kare ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Education and Science*, 33(148).

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. London: Lawrence Erlbaum Associates.

Emenogu, B. C., Falenchuck, O., & Childs, R. A. (2010). The effect of missing data treatment on Mantel-Haenszel DIF detection. *The Alberta Journal of Educational Research, 56(4),* 459-469.

Falenchuk, O., & Herbert, M. (2009). *Investigation of differential non-response as a factor affecting the results of Mantel-Haenszel DIF detection* California: American Educational Research Association.

Finch, W.H. (2011). The impact of missing data on the detection of nonuniform differential ıtem functioning. *Educational and Psychological Measurement, 71*(4) 663–683.

Garrett, P. L. (2009). *A monte carlo study investigating missing data, differential item functioning, and effect size*. Georgia State University, Unpublished doctoral dissertation.

Gelin, M.N. & Zumbo, B.D. (2003). Differential item functioning results may change depending on how an item is scored: an illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement, X(X)* DOI: 10.1177/0013164402239317.

Gierl, M.J., Jodoin, M.G., & Ackerman, T.A. (2000). *Performance of Mantel-Haenszel, Simultaneous Item Bias Test, and Logistic Regression when the proportion of DIF items is large*. American Educational Research Association.

Gonzales, A., Padilla, J.L., Dolores, H., Gomez-Benito, J., & Benitez, I. (2010). EASY-DIF: Software for analyzing differential item functioning using the Mantel-Haenszel and Standardization procedures. *Applied Psychological Measurement*. doi:10.1177/0146621610381489.

Graham, J.W. (2009). Missing Data Analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(4), 549-576.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646-675.

Gözen Çıtak, G. (2007). *Klasik test ve madde-tepki kuramlarına göre çoktan seçmeli testlerde farklı puanlama yöntemlerinin karşılaştırılması.* Doktora Tezi, Ankara Üniversitesi, Ankara

Hambletton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Harwell, M. Stone, C. A., Hsu, T.C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, 20, 101-125.

Hohensinn, C. & Kubinger K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling*, 53, 380-393.

Kan, A., Sünbül, Ö., Ömür, S. (2013). 6.- 8. sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin University Journal of the Faculty of Education*, *9*(2), 207-222.

Kothari, C.R. (2004). *Research methodology: Methods and techniques (Second Revised Edition).* New Delhi: New Age Int. Ltd.

Kristanjansonn E., R. Aylesworth, I. McDowell & B.D. Zumbo (2005). A Comparison of four methods for detecting differential item functioning in ordered response model. *Educational and Psychological Measurement. 65*(6), 935-953.

Little, R. J. A & Rubin, D. B. (1987). *Statistical analysis with missing data (2nd ed.).* New York: John Wiley & Sons, Inc.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 39, 247-264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates.

Molenberghs, G., & Kenward, M.G. (2007). *Missing data in clinical studie (1 st ed.)*. England: John Wiley&Sons.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and Simultaneous Item Bias procedures for detecting differential ıtem functioning, *Applied Psychological Measurement, 18*(4).

Osterlind, S.J. (1983). *Test item bias*. London: Sage Publication.

Padilla, J.L., Hidalgo, J.L., Benitez, I., & Gomez-Benito, J. (2012). Comparison of three software programs for evaluating DIF by means of the Mantel-Haenszel procedure; EASY DIF, DIFAS and EZDIF, *Psicologica*, 33,135-156.

Peng, C.Y.J., Harwell, M., Liou, S.M., & Ehman, L. H. (2006). *Advances in missing data methods and implications for educational research. In S. Sawilowsky* (Ed.), Greenwich: Real data analysis.

Peng, C. J., & Zhu, J. (2008). Comparison of two approaches for handling missing covariates in logistic regression. *Educational and Psychological Measurement, 68(1),* 58-77.

Pigott, T.D. (2001). A review of methods for missing data. *Educational Research and Evaluation, 7*(4); 353-383.

Robitzsch, A, & Rupp, A.A. (2009). Impact of missing data on the detection of differential item functioning the case of mantel-haenszel and logistic regression analysis. *Educational and Psychological Measurement, 69*(1): 18-34.

Rousseau, M., Bertrand, R., & Boiteau, N. (2006, April). *Impact of missing data treatment on the efficiency of DIF methods.* California: National Council on Measurement in Education.

Royce, S., Straits, B.C., & Straits, M.M. (1993). *Approaches to social research* (2nd ed.). New York: Oxford University Press.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research,* (8), 3-15.

Sedivy, S. K., Zhang, B., & Traxel, N. M. (2006). *Detection of differential item functioning with polytomous items in the presence of missing data.* California: National Council of Measurement in Education.

Selvi, H. (2013). *Klasik test ve madde tepki kuramlarına dayalı değişen madde fonksiyonu belirleme tekniklerinin farklı puanlama durumlarında incelenmesi.* Yayınlanmamış Doktora Tezi. Mersin Üniversitesi Eğitim Bilimleri Enstitüsü.

Singh, Y.K. (2006). *Fundamental of research methodology and statistics*. New Delhi: New Age Int. Ltd.

Spray, J., & Miller, T. (1994). *Identifying nonuniform DIF in polytomously scored test items.* American College Testing Research Report Series 94-1. Iowa City, IA: American College Testing Program.

Ward, W.C., & Bennett, R.E. (2012). *Construction versus choice in cognitive measurement: issues in constructed response, performance testing, and portfolio assessment.* London and New York: Routledge, Taylor & Francis Group.

Woodward, M., Smith, W.C., & Tunstall Pedoe H. (1991). Bias from missing values: Sex differences in implication of failed venepuncture for the Scottish Health Study.*Int J. Epidemiol*.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12*(3), 1-26.

Zumbo, B. D. (1999). *A Handbook on the theory and methods of Differential Item Functioning (DIF): Logistic Regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Multi-Trait Multi-Method Matrices for the Validation of Creativity and Critical Thinking Assessments for Secondary School Students in England and Greece

Ourania Maria Ventista

To cite this article: Ventista, O.M. (2018). Multi-Trait Multi-Method Matrices for the Validation of Creativity and Critical Thinking Assessments for Secondary School Students in England and Greece. *International Journal of Assessment Tools in Education*, 5(1), 15-32. DOI: 10.21449/ijate.335167

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 15-32

http://www.ijate.net

e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Multi-Trait Multi-Method Matrices for the Validation of Creativity and Critical Thinking Assessments for Secondary School Students in England and Greece

Ourania Maria Ventista *

School of Education, Durham University, Leazes Road, Durham, DH1 1TA, United Kingdom

Abstract: The aim of this paper is the validation of measurement tools which assess critical thinking and creativity as general constructs instead of subject-specific skills. Specifically, this research examined whether there is convergent and discriminant (or divergent) validity between measurement tools of creativity and critical thinking. For this purpose, the multi-trait and multi-method matrix suggested by Campbell and Fiske (1959) was used. This matrix presented the correlation of scores that students obtain in different assessments in order to reveal whether the assessments measure the same or different constructs. Specifically, the two methods used were written and oral exams, and the two traits measured were critical thinking and creativity. For the validation of the assessments, 30 secondary-school students in Greece and 21 in England completed the assessments. The sample in both countries provided similar results. The critical thinking tools demonstrated convergent validity when compared with each other and discriminant validity with the creativity assessments. Furthermore, creativity assessments which measure the same aspect of creativity demonstrated convergent validity. To conclude, this research provided indicators that critical thinking and creativity as general constructs can be measured in a valid way. However, since the sample was small, further investigation of the validation of the assessment tools with a bigger sample is recommended.

ARTICLE HISTORY

Received: 03 April 2017

Revised: 12 August 2017

Accepted: 14 August 2017

KEYWORDS

validation, creativity, critical thinking, assessment, multi-trait multi-method matrix,

1. INTRODUCTION

1.1. Research Purpose

The knowledge demands in the 21st century are not easily predictable. Therefore, the education system of each country should provide the students with skills to adapt in the needs of this changing society. It has been supported that critical thinking and creativity could address these needs (Berliner, 2011). In other words, in the 21st century there is a huge amount of knowledge available to learners. When learners are required to find solutions to their questions, they do not

*Corresponding Author E-mail: o.m.ventista@durham.ac.uk

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.335167**

have to simply recall information. Instead, they should be able to identify relevant sources and evaluate them critically. Moreover, economies and societies nowadays change rapidly, so schooling cannot prepare learners to deal with specific problems. By the time learners will finish their schooling, there will be new problems to be solved so they should be able to critically approach these issues and generate solutions creatively.

Consequently, it is not a surprise that the development of critical thinking and creativity are prioritised by school curricula across the world (for example: Australian curriculum, UK curriculum). Similarly, universities expect their students to demonstrate critical and creative thinking and include these skills in their scoring rubrics. Therefore, critical thinking and creativity are judged to be crucially important within educational systems.

Despite their growing importance, the measurement tools of creativity and critical thinking as generic skills are not well established in primary and secondary education. As a result, when primary and secondary school students are assessed, traditional forms of assessment, which focus mainly on attainment, are used.

Hence, this paper investigates to what extent assessments which measure creativity and critical thinking as general constructs can be reliable and valid. To be more precise, concerning reliability, this paper focuses on the internal consistency of the measurement tools. For validity, this paper examines the discriminant (or divergent) and convergent validity. These are important elements to be investigated since there is no sufficient evidence for these psychometric properties. Although there is recent research which examines the relationship of students' performance between sub-sections of Torrance test (Yoon, 2017) or team creativity (Jiang & Zhang, 2014), there is a lack of studies which examine and establish the convergent validity among creativity tests (Plucker & Maker, 2010; Yoon, 2017).

Similarly, for critical thinking there are examples of studies attempting the validation of critical thinking as a subject-specific skill (Tiruneh et al., 2017). However, there is no evidence about the convergent validity between measurement tools of critical thinking.

Even when convergent validity of critical thinking measurement tools is examined, it is not established on comparison of performances in critical thinking assessments. For instance, recently a critical thinking tool for primary school students was developed. The researchers attempted to establish the criterion validity (which is a type of convergent validity) by comparing the performance of students with their grades of students in arts, instead of another critical thinking assessment (Gelerstein et al., 2016). This means that convergent validity was considered, but not in the most rigorous way.

Consequently, there is not sufficient evidence of the validation of creativity and critical thinking measurement tools. Hence, this research contributes to this area and discusses psychometric properties of assessments of creativity and critical thinking. For the purpose of this article, first, the constructs of critical thinking and creativity are defined and operationalised, then, the processes that the validation of measurement tools achieved are discussed. Next, the research methodology is presented, and, finally, the results of this research and its limitations are reported.

1.2. Defining the constructs

Creativity and critical thinking are the focal points of this research. Both terms can be perceived in different ways, but it is fundamental for both constructs to be defined before deciding on their assessments. Critical thinking 'is the intellectually disciplined process of active and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information

gathered from, or generated by, observation, experience, reflection, reasoning, or communication' (The Critical Thinking Community, 2013). According to Ennis (1993), critical thinking involves judging arguments and the credibility or sources, identifying conclusions and assumptions and drawing warranted conclusions. While Ennis (1993) defines "critical thinking as a reasonable reflective thinking that is focused on deciding what to believe and do", Lipman (1987) explains that the use of the word 'reasonable' can lead to circularity and criticised this definition as restrictive. According to Lipman (1987), critical thinking is employed for numerous other aims and does not always lead to a clear-cut conclusion. Lipman (2003) postulates that critical thinking is based on criteria, is self-corrective and sensitive to context. A further definition of critical thinking supports that it involves six basic cognitive aspects: interpretation, analysis, evaluation, inference, explanation and self-regulation (Facione, 1990, 2015). For this research, the working definition of critical thinking consists of observation, analysis, synthesis, evaluation and interpretation of arguments within specific contexts.

Creativity is perceived as a broad term which includes other sub-characteristics such as divergent thinking, convergent thinking, openness to explore new ideas and listening to "inner voice" (Treffinger, Young, Selby, & Shepardson, 2002). According to this paradigm, creativity includes critical thinking. Guilford (1967) supports that problem-solving is the same phenomenon as creative thinking. In order for something to be perceived as creative, it should have two main characteristics: to be original and useful (Rungo & Jaeger, 2012). According to the definition of the National Advisory Committee on Creative and Cultural Education (1999), however, creativity has four - instead of just two - typical characteristics: imagination, purposefulness, originality and a new product with merit. Similarly, Mednick (1962) defines creative thinking as the procedure through which associated components are combined in a new way and this combination is a useful one. In recent years many researchers have accepted the standard definitions of creativity (Weisberg, 2015). By examining studies regarding the definitions of creativity (Kampylis & Valtanen, 2010), it can be concluded that most of the recent definitions involve trivial additions or syntheses of previous ones. Weisberg (2015), however, questions the inclusion of "value" in the definition of creativity, since its evaluation appears to be too subjective and unreliable. As a result, for the purposes of this research creativity is operationalised as a combination of fluency, innovation, novelty and imagination.

1.3. Validation

Having discussed the working definitions of the two main constructs, issues regarding validation of assessment tools are discussed. This paper investigates to what extent critical thinking and creativity assessments can be considered valid. The first issue to be discussed is whether the validity is a psychometric property of a test or a characteristic of the interpretation of the test. On the one hand, it has been supported that a test is valid when it measures what is supposed to measure, so the validity is a psychometric property of the test. On the other hand, it has been supported that the interpretation is the one which can be valid or invalid and a test cannot be itself valid or invalid. This means that a test can be valid for one interpretation, but invalid for another one (Coe, 2012; Newton, 2012).

The second issue concerns the ways that validation can be achieved. Five sources of evidence can support the validation process; test content, response processes, internal structure, relations to other variables and consequences of testing (Sireci, 2009, p. 30). Specifically, about the test content, Kane (2009) states that if the task of a test is close to the performance of interest then there is no need for strong evidence for the content of the test for it to be valid.

With reference to the internal structure as a process of validation, the factors included in a test are considered. This research used Cronbach's Alpha as an indicator of internal structure. Although the relations to other variables is usually called criterion validity, in critical thinking and creativity assessments, there is not a widely accepted gold standard to be considered as criterion. Instead, this research used what Campbell and Fiske discuss (1959) as a validation method: convergent and discriminant validity. Messick (1995) also mentions this method as one aspect of validity, which is related to the external evidence for the quality of an assessment. Convergent validity exists when results from measures that measure the same construct are correlated, while discriminant validaty when the scores of tests which measure different constructs do not correlate. Particularly, convergent validity was sought between the measurement tools which measured the same construct (either creativity or critical thinking) and divergent validity between the measurement tools which measured different constructs (critical thinking and creativity). This implies that this research accepts that critical thinking and creativity are not the same constructs, even though some researchers might have expressed the opinion that they are both part of productive thinking (Facione, 2015; Newton, 2014).

2. METHODOLOGY

2.1. Method

For the selected validation process, collection of data was required. In this case, data was the scores in the assessments. This paper presents the results of research conducted in Greece and its replication in England. As previously mentioned, the validation of the measurement tools attempted to be done with using the multi-trait multi-method matrices (Campbell & Fiske, 1959). This analysis requires the use of at least two traits and two methods. The two traits were creativity and critical thinking and the two methods were written and oral assessments.

As multi-trait multi-method matrices were used, emphasis was put on convergent and discriminant validity. So the hypothesis was that if tests of critical thinking indeed measured critical thinking then the scores that students achieved in both critical thinking tests would be correlated with each other (convergent validity). On the other hand, their critical thinking scores would be less or not correlated with measurements of creativity (discriminant validity), since the assessments measured different constructs. With the exact same logic, there was a similar hypothesis for the creativity measurement tools. If the creativity scores were valid and measured what they supposed to measure, then the scores that the students would achieve in creativity assessments would correlate with each other (convergent validity) and would not correlate with their performance in critical thinking (discriminant validity).

Lastly, because the methodology required correlating scores of the tests, it has to be clarified that there is no lower limit for the sample size when conducting a correlation study. The sample size, however, affects the confidence intervals for the correlation. With small sample sizes, even a slight increase in the number of participants significantly reduces the length of confidence intervals. However, it has been supported that when increasing the number of participants to more than 24 participants, there is a loss of sample size impact on the length of the confidence intervals (Johanson & Brooks, 2010, p. 397). Finally, it has to be mentioned that the recommended number of participants for pilot studies is usually around 30 (Johanson & Brooks, 2010).

2.2. Replication

Seven months later the research was replicated in a secondary school in the North East of England. The purpose of this replication was not the direct comparison of the two countries but to increase the sample size. In Greece, there were only 30 students, so it was judged appropriate to collect some additional data. However, it was interesting to investigate whether the previous results would be also found in a new situation. Moreover, replication was conducted specifically in England in order to exclude the possibility of effects of translation issues, which might have affected the Greek sample.

The results of each study are presented separately because there was one small change in the methodology and because the data collection took place at different times. As I am not a native English speaker, my accent could contribute to a construct irrelevance in the oral assessment of critical thinking. For this reason, students were given three different options than the Greek students. The Greek students had a text read to them, while the English students could choose between the researcher reading the text or them reading it aloud or silently. There is the assumption that they chose wisely in order to maximize their performance in the test and indirectly minimize the potential construct irrelevance.

Even though it would have been preferable to keep the conditions exactly the same as in Greece, it was not possible. Instead of giving them this choice, the alternative of having a recording of the letter read by a native speaker was considered. However, this was too impersonal and could have not taken into consideration the conditions in the room. Hence, it was judged as a bigger change in the methodology compared to allowing the student to choose their preferred method of accessing the text.

2.3. Participants

The initial research took place in a secondary school in Greece with 30 participants aged 13-15 years old. Students of these ages were targeted because there are more available assessment tools for these ages compared to primary school students. The specific school was selected based on the willingness of the headteacher to provide time and space for the research needs. The school was in a suburban area of northern Greece. The students were randomly chosen by the class lists. No student refused to participate and there was no attrition.

In the replication study, the sample was 21 twelve-year old boys who were students in a secondary school. It was not possible to gain access to older students as in the Greek sample. However, the tests were age-appropriate. In this sample 4 participants refused to narrate a fairy tale and this research believes that they felt uncomfortable to do so. British Education Research Association (BERA) guidelines stipulate that participants can withdraw at any point. During the research and during the replication of the research two of the students withdrew (BERA, 2011) .

2.4. Ethics

Before conducting both studies, ethical approval was obtained by the School of Education Ethics Committee at Durham University. Both of the studies followed the BERA guidelines (2011).

3. ASSESSMENT TOOLS

3.1. Critical thinking

The tools used for the critical thinking in the written method were a combination of the deduction items of the Cornell Reasoning Test (Ennis et al., 1964) and items based on the test of appraising observation (Norris & King, 1984). The reasoning test provides "if" statements to the students who should judge whether the last sentence would be a warranted conclusion by deductive reasoning. A choice of "maybe" is also given to the students in this test, as in some cases the data are insufficient for them to decide. The test of appraising observation narrates two stories to the students. Each item of the test provides two statements to the students. The students should judge which of the two statements is more believable. In order to judge effectively, the students should also consider the context of the two stories as a factor.

The time given for these tests was one hour and due to this time limitation only a few items were used. Both tests are quite extensive and, thus, since the aim was not to examine the reliability and validity of the specific existing tools, but to examine whether it was possible to measure critical thinking as a general construct, only a few questions of each test were used. In order to improve the internal consistency of the initial tests, similar questions appear multiple times. In this research, fewer questions were chosen. The questions were judged appropriate and sufficient to operationalise the construct of critical thinking as defined by this research.

Additionally, both of the tests are age appropriate. The Cornell Test Level X (Ennis, Gardiner, Guzzetta, Morrow, Paulus & Ringel, 1964) was deemed appropriate for secondary school students and used in previous studies for evaluating critical thinking in students of this age or even a little older (Iozzi & Cheu, 1978). The last version of appraising observation test is also suitable to assess secondary school students (Norris & King, 1984).

The critical thinking tool used for the oral assessment of critical thinking was based on an established tool (Ennis & Weir, 1985) suitable to test sixth grade to university students. During this assessment, the students were requested to judge presented arguments. The researcher first articulated the main purpose of the letter - the author tried to persuade the listener of the benefits of the prohibition of overnight parking- and then read the letter. The researcher elucidated that students should take a position and either be persuaded or not by the argument in each paragraph to justify their position and share any thought related to the paragraph. The reason why the letter was read by the researcher to the Greek students was to exclude construct irrelevance. It has been supported that the reading ability in tests can play an important role (Hewitt & Homan, 2003). Reading ability is irrelevant to critical thinking and should not be embodied in critical thinking assessments. The oral assessment did not disadvantage students who have reading difficulty. They could also ask for clarification for words that they didn't understand. They had sight of a printed version so as not to disadvantage students who were not used to listening to texts.

3.2. Creativity

For the written assessment of creativity a combination of tests was used (Getzels & Jackson, 1962). Firstly, students had to think as many possible uses for common objects, such as a brick. Secondly, students were given partially complete images and instructed to complete them by drawing around them to illustrate what they imagined the images were. An activity similar to the latter can also be found in the Torrance Test of Creative Thinking (Torrance, Ball & Safter, 2008). The number of responses given by the students and the degree of originality of their responses were assessed.

For the oral assessments of creativity the students were asked to narrate a fairy tale. For the fairy tale a scoring rubric was created. The rubric evaluated the content of the students' stories by combining indicators of imagination. These indicators were the number of mentioned typical elements found in fairy tales, referred to as *functions* (Propp, 1968), the presence of creative characteristics that can be in fairy tales (Rodari, 1996) and the presence of humour and violence in the story. The latter two characteristics are usually connected with creativity (Getzels & Jackson, 1962; Nusbaum, Silvia & Beaty, 2017).

The oral assessment resembled a real-life task with a specific purpose as the communicative language approach would suggest (Richards, 2005). Participants were presented with a real life situation: "*A younger cousin or a sibling of yours has just asked you to narrate a fairy tale. I will give you three minutes to think about the fairy tale you are going to narrate and about this time again to narrate it*". The choice of the activity was grounded in results of prior research investigating gender and ethnicity differences in creativity. Even though males had the self-perception of being more creative on science-analytic and sports tasks and females more on social-communications and visual-artistic tasks, both genders were equally assumed to be creative in verbal-artistic activities (Kaufman, 2006). For this reason a type of verbal activity was set. Nonetheless, it is accepted that for the previous finding, since it is based on self-reported questionnaires there may be a gap between perceived creative strengths and actions, and also that the respondents' opinions and beliefs may not be stable (Foddy, 1993).

3.3. Norm-referenced tests

The two written tests of creativity were norm-referenced measurements because there was a comparison between the performances of the students (Cox & Vargas, 1966). The score of unique answers attributed to the students related to the other participants' responses. Thus, an answer was characterised unique only if no other participant had mentioned this particular answer. Silvia (2015) highlights the significance of this flaw in the creativity tests; the uniqueness grade is sample-dependent. In other words, as the sample increases, the likelihood of a unique answer decreases.

To ameliorate this, the researchers could pre-decide the size of group. For example, the sample for this test could always be 30 students and each reply could be judged unique when it has not been mentioned by the particular number of students. It is accepted that this could not provide a solution for the problem of a student having high performance in a less creative group and be judged to have average performance when compared to a more creative group. Nevertheless, sample-dependence cannot be completely avoided in the norm-referenced tests.

3.4. Matching the assessments to the construct definitions

It is important to discuss the tools used for this research in relation to the aspects of the constructs measured. The appraising observations test assessed the ability of the students to evaluate which statement is more believable. Analyzing and synthesizing can also be assessed by the test (Treffinger et al., 2002). The reasoning test evaluated deductive reasoning. The Ennis & Weir letter (1985) required evaluation of specific arguments. Therefore, these assessments fit the aforementioned definition of critical thinking.

The 'test of different uses for tools' and the 'pattern meanings test' (Getzels & Jackson, 1962) did not have a single correct answer. The only variables measured in this test were originality (how many answers are unique between the answers of all the participants) and fluency (the number of answers mentioned) firstly at the suggestion of the test author (Getzels & Jackson, 1962)

and secondly because these variables can be measured objectively. Concerning the narration of the fairy tale, it mainly attempted to evaluate imagination and innovation, which are characteristics of the creativity (El-murad & West, 2004). Sense of humor as a characteristic of openness was assessed by the oral assessment of creativity. Consequently, creativity assessment also fit the working definition of creativity adopted by this research.

3.5. Translation and adjustment of the Tools in Greek

Measurement instruments were cautiously translated in the Greek language using the back-translation method (Su & Parham, 2002). Furthermore, for the oral assessment of creativity, the content was also slightly adjusted. The town took the name of the town in which the test was administrated, road names were taken from roads in the town and also the name of the authorities 'Director of the National Traffic Safety Council' and the 'National Association of Police Chiefs' were replaced with the respective Greek terms. This aimed to provide the students with a purpose and a motivation to read the test (Richards, 2005).

4. RESULTS AND DISCUSSION (Study in Greece)

The tools are going to be discussed according to their reliability and validity. There are different types of reliability and validity. For the purpose of this research, the reliability is discussed as internal consistency and validity as convergent and discriminant validity.

Table 1. Multi-trait multi-method matrix (Greece)

		WRITTEN TESTS Method 1			ORAL ASSESSMENT Method 2	
		Critical thinking	Creativity: DUO	Creativity: PM	Critical thinking	Creativity
Written tests Method 1	Critical thinking: only reasoning	0.758				
	Creativity: Different Uses of Objects	-0.021	0.817			
	Creativity: Pattern Meanings	-0.376 *	0.719**	0.925		
Oral Assessment Method 2	Critical thinking	0.199	0.139	0.216	0.483	
	Creativity	-0.299	-0.010	0.169	0.257	0.743

* $p < 0.5$ (statistical significance)
** $p < 0.1$ (statistical significance)
Light blue: the cells which show just the internal consistency of the measurement tool
Light green: the cells which show correlation between monomethod and the same trait.
Light pink: the cells which show correlations between heterotrait and monomethod cells (creativity or critical thinking compared with each other and assessed by the same method).
Purple: the cells which show correlations between heterotrait - heteromethod cells.
Orange: the cells which show correlations between monotrait - heteromethod cells.

4.1. Internal Consistency of the Measurement Tools

To consider the reliability of the measurement tools, internal consistency was examined and Cronbach's Alpha was used as an indicator of internal consistency. Cronbach's Alpha should not be used as proof of all types of reliability. It is only related to the correlation of the items and it is the 'mean of all split-half reliabilities for a given test application' (Johnson & Johnson, 2009, p. 14). The internal consistency of the items based on the appraising observation tests was low and it could not be improved even by deleting some items. Thus, these items were excluded by the matrix.

Some of the reasoning items were found to have negative correlation so they were deleted. An item that has negative correlation tends to be answered incorrectly by otherwise high scoring students. One of those items had negative stem. Negative statements in the stem should be avoided (Haladyna, 1994) because it may cause confusion. Two items at the end of the test also had negative correlation, but these items did not seem to differ from the other items. The fact that they were towards the end of the test may be the cause of those items having negative correlation. The students may have been tired or bored by the end of the test.

The results for the reasoning items in the written assessment of creativity had indicated strong internal consistency (a = 0.76). The creativity assessments for the written method also had high reliability (a = 0.81 and a = 0.92), which is comparable with alpha scores required for high-stakes assessment. The oral assessment of creativity had also high internal consistency (a = 0.74). Consequently, even though critical thinking and creativity are multi-facet constructs, when the tests are focused on particular aspects, such as only reasoning or imagination, then high internal consistency can be expected.

The oral assessment of critical thinking was found to have moderate internal consistency (a= 0.48) which could have been a consequence of the test having a few items. With more items, the reliability of the test may have been higher, however, the increase of the number of the items cannot be assumed to substantially increase of the quality of the test even if this is a way to increase internal consistency. For example, by asking similar questions the length of the assessment and Cronabach's alpha increases. However, the quality of assessment remains the same. The low alpha might be explained by the fact that the test was not a multiple-choice test. Multiple choice items are usually preferred in tests because they increase reliability, but this does not mean that they secure the validity of the tests (Burton, Sudweeks, Merrill & Wood, 1991; Lambert & Lines, 2000). Thus, even though the oral assessment had lower internal consistency than the other assessments, it might have been a more valid method of testing critical thinking. Even though there are researchers who support that there cannot be valid inferences without reliability (Koretz, 2006), there are others who advocate that if reliability is perceived merely as consistency among measures then validity may be without reliability (Moss, 1994). Moss (1994) supports that less standardised forms of assessment may be valid without being reliable and 'as assessment becomes less standardised, distinctions between reliability and validity blur' (p.7).

4.2. Convergent and Discriminant Validity

The multi-trait and multi-method matrix presents the convergent and discriminant validity between the measurement tools (Table 1). The written test of critical thinking was validated based on convergent and discriminant validity. Specifically, it was correlated with the oral assessment measuring critical thinking (convergent validity), but not correlated with the creativity assessments (discriminant validity).

The written test of critical thinking had discriminant validity with the three creativity tests (r = - 0.02, r = -0.38 and r = - 0.3). This means that there was not a linear relationship which links the performance in the reasoning items with the performance in the creativity tests of fluency, innovation and imagination. As a result, the reasoning test measured something different from the creativity tests.

The performance of students in the reasoning items had a very weak linear relationship with their performance in the oral assessment of creativity (r = 0.2). This means that the two assessment had, to some extent, convergent validity, but without strong evidence. The low correlation between the scores in the two assessments of critical thinking can be explained because the two tools evaluated different aspects of critical thinking. The written test was focused on deductive reasoning, while the oral assessment on the argument evaluation within a specific context.

The scores of the oral assessment of critical thinking was correlated equally with those of the oral assessment of creativity (r = 0.14 and 0.22) and the written test of evaluating critical thinking (r = 0.2). Similarly, the scores of the oral assessment of creativity was more correlated with the scores of the oral assessment of critical thinking (r = 0.26) rather than those of the creativity assessments (r = - 0.1 and r = 0.17). Thus, the performance of the students in the oral assessments correlated more with each other than with their performance in tests which evaluate the same constructs with different methods. This is not a surprising finding. Paradoxically it is common to identify higher correlation between the scores of heterotrait and homomethod assessments, rather than the homotrait and heteromethod (Coe, 2012).

Furthermore, in this case, slight correlation between the scores that students achieved in critical thinking and creativity assessments is expected, because creativity and critical thinking - as they have already been defined - can be related to each other and be perceived as sub-categories of productive thinking (Newton, 2014).

The scores of the two written assessments of creativity were highly correlated with each other with a strong linear relationship (r = 0.72). In other words, the students who scored highly in the one test also scored highly in the other test, and the students who scored low in one, they also scored low in the other test. This suggests that both tests measured the same thing and that evidence of convergent validity was strong.

This last finding can be considered a positive indicator for future assessment of creativity. For these two tests, it is possible that there is concurrent validity, as they both also have independently high reliability (Lambert & Lines, 2000). Both tests evaluated mainly the same elements of the creativity construct, fluency and innovation by using the same method. The high correlation between their scores demonstrates that as long as the same side of a multifaceted construct is evaluated with the same method using two different assessments, convergent validity between these assessments can be expected.

What requires explanation is the fact that the scores of the two written assessments of creativity were poorly correlated both with those of the critical thinking oral assessment (r = 0.14 and r = 0.22) and with the creativity oral assessment (r = - 0.01 and r = 0.17). More specifically, the low correlation between the written assessments of creativity and the oral assessment of critical thinking can be explained if the two constructs are considered elements of the general construct productive thinking.

The low correlation between the written assessments and the oral assessment of creativity (r = - 0.1 and r = 0.17) can be used as a lucid demonstration that creativity is a multi-faceted concept

and the assessments evaluate different aspects of the same construct. The written test about the use of objects measured fluency and innovation, while the oral assessment measured verbal imagination. Thus, students might have been creative in some aspects, but not in others. In other words, different measurements tools of creativity using different methods were not found to be highly correlated. This finding is line with studies in creativity literature which suggested that people might perform differently in different tasks which require creativity (Hocevar, 1979).

To summarise, convergent and divergent validity were found for the written critical thinking assessment. Similarly, the creativity assessments had high convergent validity only when the same method and the same facets of the construct were assessed. The research in Greece revealed some positive indicators for the evaluation of critical thinking and creativity as general constructs.

5. RESULTS AND DISCUSSION (Replication Study in England)

A few months later the study was replicated in England. The results observed were similar to those derived from the Greek sample.

5.1. Internal Consistency of the Measurement Tools

When the research was replicated, the internal consistency of the measurement tools was also found to be relatively high. The reasoning items in the written assessment were found with similar internal consistency values as in Greece (a= 0.74). All the assessments of creativity had high alpha scores (a = 0.8), similar to the Greek sample data. These values of internal consistency are sufficient to enable the assessments to be used as high-stakes. The high internal consistency values could be explained by the fact that all the three creativity assessments measure a narrow and specific aspect of creativity.

Concerning its internal consistency, the data relating to the questions based on the appraising observation test indicated a low alpha score when implemented in Greece, but with the English sample it was slightly higher (a=0.52). For a multiple-choice test to have such a low alpha score is concerning as it contradicts with the usual expectation of multiple-choice items to be more reliable assessments (Burton et al., 1991).

Finally, the oral assessment of critical thinking had a higher internal consistency (a = 0.57) than the Greek sample. The test was not a multiple-choice test and this might affect its internal consistency.

5.2. Convergent and Discriminant Validity

When replicating the research in England (Table 2) the evidence was similar to the results from the Greek data (Table 1), as the multi-trait multi-method matrices suggested. The written assessment of critical thinking was also validated with convergent and discriminant validity, as with the Greek sample. The evidence for convergent validity in the English sample was stronger than the Greek one, since a moderate linear relationship between the written assessment and oral assessment of critical thinking was found (r=0.44). This relationship suggested that the students who scored highly in one test usually tended to score highly in the other test as well. The relationship between the two tests was much stronger compared to what was found in the Greek sample (r = 0.2). A possible explanation might be an issue of translation or cultural differences in the critical thinking tests in the Greek sample.

Table 2. Multi-trait multi-method matrix (England)

		WRITTEN TESTS Method 1			ORAL ASSESSMENT Method 2	
		Critical thinking	Creativity: DUO	Creativity: PM	Critical thinking	Creativity
Written tests Method 1	Critical thinking: reasoning items	0.741				
	Creativity: Different Uses of Objects	0.251	0.813			
	Creativity: Pattern Meanings	0.208	0.477*	0.879		
Oral Assessment Method 2	Critical thinking	0.437	-0.357	-0.383	0.566	
	Creativity	-0.040	0.159	0.228	-0.332	0.845

* p < 0.5 (statistical significance)
** p < 0.1 (statistical significance)
Light blue: the cells which show just the internal consistency of the measurement tool
Light green: the cells which show correlation between monomethod and the same trait.
Light pink: the cells which show correlations between heterotrait and monomethod cells (creativity or critical thinking compared with each other and assessed by the same method).
Purple: the cells which show correlations between heterotrait - heteromethod cells.
Orange: the cells which show correlations between monotrait - heteromethod cells.

For the written test of critical thinking there was a very weak relationship with the written tests of creativity (r = 0.25 and r = 0.2), but no relationship with the oral assessment of creativity (r = - 0.04). The first two assessments might be slightly correlated because they use the same method (written) as the reasoning items and it has been found that there is correlation between assessments which use the same method independently of the construct (Coe, 2012). However, the lack of relationship between the reasoning items and the oral assessment of creativity established the discriminant validity between the assessments.

Moreover, discriminant validity between the oral assessment of critical thinking and creativity measurement tools was reported (r = -0.36, r = - 0. 38 and r = - 0.33). Therefore, the data from the English sample validated the critical thinking tools with both convergent and discriminant validity.

The scores of the two written creativity tests were found with a sufficient linear relationship to establish convergent validity both in Greece (r= 0.72) and in England (r = 0.48). Thus, as the same side of a multifaceted construct is evaluated and the same method is used, correlation between the tests can be expected.

The results of the two written assessments of creativity were found almost equally correlated with the written assessment of critical thinking (r = 0.25 and r = 0.25) and the oral assessment of creativity (r = 0.16 and r = 0.23). However, as mentioned previously, there are examples of studies

which demonstrate that the method by which students are assessed sometimes plays a more crucial role than the construct on which they are assessed (Coe, 2012).

With reference to the oral assessment of creativity, there was validation of the assessment. Convergent validity was found between the oral assessment of creativity and the two tests of creativity (r = 0.16 and r =0.23). The convergent validity, however, was not supported by high correlation between the creativity assessments. This is expected, because the oral assessment of creativity did not examine the same aspects of creativity concept as the written assessment of creativity. This finding confirmed that creativity characteristics vary within a person and no person can have all the creative characteristics (Treffinger et al., 2002). In multi-faceted constructs like creativity, convergent validity can be sought between assessments which evaluate the same aspects of the construct.

Furthermore, discriminant validity was found since the oral assessment of creativity was not correlated with the two critical thinking assessments (r = - 0.04 and r = - 0.33). The lack of correlation between the performances of the students in the oral assessment of creativity and the critical thinking tests suggested that they measure different concepts. Therefore, there was discriminant validity which also supported the validation of the measurement tools of creativity and critical thinking.

To conclude, the assessments in the multi-trait and multi-method matrix in England were found to be valid concerning their convergent validity and discriminant validity. Consequently, the replication of the study confirmed the findings of the initial study in Greece and supported with even stronger evidence that critical thinking and creativity can be evaluated as general constructs in a valid way.

5.3. Is critical thinking and creativity culture and knowledge dependent?

As it has been previously said, the purpose of collecting data from two different countries was not their comparison. Besides, the sample was too small to enable such a comparison. However, by replicating this study in two different schools in two different countries and by perceiving critical thinking and creativity as general constructs and not subject-specific, it is reasonable to question to what extent the performance of the students was culture and knowledge dependent. For a deeper understanding of potential differences, there was an examination of the recorded material of the oral assessments. This material gave access to the students' thinking process. In the narration of the fairy tale no significant cultural differences were identified. The themes that emerged in the students' stories were similar. Moreover, this task did not demand any knowledge and thus knowledge did not appear to affect the performance of the students.

This was not the case with the relationship between knowledge and the evaluation of arguments in critical thinking assessment. Some students were not critical because of the lack of specific knowledge. Particularly, students were persuaded by an argument presenting results of a one-day experiment. Being students in a secondary school and without research knowledge they could not realise that results of one day experiment could not support generalisation. Therefore, sometimes prior knowledge is required to be critical. This is in agreement with the ideas of some of academics. For example, McPeck (1981, 1990) supports that critical thinking is subject-specific and in order for somebody to be critical they should have knowledge of the topic. This stance opposes Ennis' whose definition and assessments have been broadly accepted by this research. However, it should be recognised that it is valid to evaluate critical thinking as a high-order thinking skill of a subject as the Bloom's taxonomy would espouse (Krathwohl, 2002), when there

are also knowledge requirements in the assessment. Nevertheless, as the findings of this research suggested, critical thinking tests which do not require prior knowledge can be constructed.

No cultural differences were identified when the critical thinking performance of students in England and Greece were compared. However, when one of the arguments in the oral assessment of critical thinking discussed driving to work during rush hour, three students in Greece suggested arriving to work slightly late in order to avoid rush hour traffic. This was not suggested by English students. The sample was too small to lead to generalisation, but this might suggest some cultural differences. Hence, critical thinking assessments could be biased because of cultural differences.

Finally, the arguments used in the oral assessment of critical thinking were adjusted in the Greek language and context by also using a town familiar to the students. This adjustment aimed to make the context more realistic and motivate some students. However, it confused other students who became fixed on the real traffic problems of that specific town. Therefore, if the topic in the critical thinking test is relevant to the daily life of the students, this may affect their judgment. The students might adhere to the specific stimulus provided, which could restrict their judgment. This is in line with what Lipman (2003) supported; critical thinking is -and should be - related to the context.

6. LIMITATIONS

The two matrices in this research can only provide positive indicators for the validation of the tools, because the research design had several limitations. Specifically, the sampling method and the small number of participants do not allow generalisation of the conclusions about the effectiveness of the assessment tools. However, the assessments were conducted by only one researcher and it was infeasible to conduct more oral assessments (each of them lasted approximately 30 minutes). It is suggested that future studies use a bigger sample.

Additionally, the tests had no consequences for the students, and their motive to complete them was not examined. They may have merely guessed several of the questions as there were no aftereffects. What is more, narrating a fairy tale may inadequately motivate teenagers, especially boys. Some teenagers may feel in an inconvenient position when someone asks them to narrate a fairy tale. Moreover, with solely one rater, interrater reliability could not be examined. In the oral assessment halo effects may have been present to some extent which may have influenced marking (Nisbett & Wilson, 1977). Finally, the tests were translated for implementation in Greece. Even though back-translation took place, translation may still affect the results (Su & Parham, 2002).

For future researchers the replication of the research with a bigger sample is recommended. In both matrices, the creativity tool ‘narrating a fairy tale’ used in the oral assessment found highly reliable but not particularly correlated with any other test. This might be either because it evaluates different aspects of creativity or because the gender or the age of the students influenced their motivation and involvement in this task. In future research, it would be useful to pilot this tool with students in primary school and attempt to examine the convergent validity with other established creativity tests which evaluate the same aspect of creativity. Moreover, it is crucial for the convergent validity of this test with linguistic ability tests to be examined. It might be the case that this tool has high construct irrelevance by including general language ability since participants have to express their thoughts and tell a story by not only demonstrating an isolated creativity skill.

7. CONCLUSIONS

Critical thinking and creativity as general constructs can be measured. Most of the assessments had moderate or high internal consistency. Furthermore, internal consistency was found to be independent of the format of the tests, as one of the multiple-choice assessments was found to be the least reliable.

By using convergent and discriminant validity for the tools' validation, there was some evidence that critical thinking and creativity tools which evaluate these constructs as general can be valid. Discriminant validity between critical thinking and creativity tools was identified in almost all of the instances in both countries' data matrices.

The value of convergent validity between the assessments which measure the same constructs in some of the cases has been low. However, this finding is justifiable because in some cases even though both tests measured the same construct, they measured different aspects of the same construct. Hence, if creativity and critical thinking are to be evaluated, the convergent validity of the tests should be sought between tests which assess common sides of the construct. The validation of the tools could not be achieved when the assessment tools measured different sides of the same construct.

In a few cases, assessments using the same method were found highly correlated to each other even though they measured different constructs. This suggests that the assessment method can play a crucial role in the students' performance in the thinking skills assessments.

As a final remark, since critical thinking and creativity are multi-faceted constructs, multi-assessment is recommended, because students might perform well in an assessment which measures one of the facets, but not in another which measures one of the other facets.

Disclosure statement

No potential conflict of interest was reported by the author.

Acknowledgements

I would like to thank Prof Rob Coe for his support in all the aspects of this research (both methodological and administrative). I would also like to thank Prof Stephen Gorard for his insightful comments and the four reviewers of the IJATE for their detailed and constructive feedback on my manuscript. Finally, I would like to thank Miss Sinead Flinders for her invaluable help, which significantly improved my manuscript.

8. REFERENCES

Australian Curriculum (n.d.) *Critical and Creative Thinking*. Available at: https://www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/critical-and-creative-thinking/ (access: 6 August 2017)

BERA (2011). *Ethical guidelines for educational research*. British Educational Research Association. Available at BERA website: https://www.bera.ac.uk/researchers-resources/publications/ethical-guidelines-for-educational-research-2011 (access: 5 August 2017)

Berliner, D. C. (2011). 'The Context for Interpreting PISA Results in the USA: Negativism, Chauvinism, Misunderstanding, and the Potential to Distort the Educational Systems of Nations'. In Pereyra, M.A., Kotthoff, H. & Cowen, R. (ed.) PISA Under Examination:

Changing Knowledge, Changing Tests, and Changing Schools. (pp. 77-96). Rotterdam: Sense Publishers

Burton, S.J., Sudweeks, R.R., Merrill, P.G. & Wood, B. (1991). *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty*. Brigham Young University Testing Services and The Department of Instructional Science.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and Discriminant Validation by the Multitrait-multimethod matrix. *Psychological Bulletin*, 56 (2), 81-105

Coe, R. (2012). 'Conducting Your Research: Inference and Interpretation'. In Arthur, J., Waring, M., Coe, R. & Hedges. L.V. (ed.) Education Research: Methods and Methodologies. (pp. 41-52). London: Sage

Cox, R.C. & Vargas, J.S. (1966). A comparison of Item Selection Techniques for Norm-Referenced and Criterion-Referenced Tests. University of Pittsburgh.

Critical Thinking Society (2013). *Defining Critical Thinking.* Available at: http://www.criticalthinking.org/pages/defining-critical-thinking/766 (Accessed: 28 January 2015).

Department for Education (n.d.) *National Curriculum in England: Framework for key stages 1 to 4.* Available at Gov.UK website: https://www.gov.uk/government/publications/national-curriculum-in-england-framework-for-key-stages-1-to-4/the-national-curriculum-in-england-framework-for-key-stages-1-to-4#the-school-curriculum-in-england (access: 6 August 2017)

El-Murad, J. & West, D. C. (2004). The Definition and Measurement of Creativity: What do we know?. *Journal of Advertising Research*, 44(2), 188-201. doi: 10.1017/S0021849904040097

Ennis, R.H. (1993). Critical thinking assessment, *Theory Into Practice*, 32(3), 179-186. doi: 10.1080/00405849309543594

Ennis, R.H., Gardiner, W., Guzzetta, J., Morrow, R., Paulus, D. & Ringel, L. (1964). *Cornell Critical Thinking Test Series. The Cornell Critical Reasoning Test. Form X*. University of Illinois.

Ennis, R.H. & Weir, E. (1985). *The Ennis-Weir Critical Thinking Test: Test, Manual, Criteria, Scoring Sheet. An instrument for teaching and testing*. Pacific Grove: Midwest Publications

Facione, P. A. (1990). *Critical Thinking: A statement of expert consensus for Purposes of Educational Assessment and Instruction (The Delphi Report).* California State University.

Facione, P.A. (2015). *Critical Thinking: What it is and why it counts.* Revised. Insight Assessment.

Foddy, W. (1993). *Constructing Questions for Interviews and Questionnaires.* Cambridge: Cambridge University Press

Gelerstein, D., del Río, R., Nussbaum, M., Chiuminatto, P., & López, X. (2016). Designing and implementing a test for measuring critical thinking in primary school. *Thinking Skills and Creativity*, 20, 40-49.

Getzels, J.W. & Jackson, P. W. (1962). *Creativity and Intelligence: Explorations with Gifted Students*. London and New York: John Wiley and Sons Inc.

Guilford, J.P. (1967). *The nature of Human Intelligence.* New York: Mc Graw-Hill Book Company

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. UK: Lawrence Erlbaum Associates.

Hewitt, M.A. & Homan, S.P. (2003). Readability level of standardized test items and student performance: The forgotten validity variable, *Reading Research and Instruction*, 43(2), 1-16.

Hocevar, D. (1979). Measurement of Creativity: Review and Critique. *Annual Meeting of the Rocky Mountain Psychological Association*, Colorado, 12-14th April

Iozzi, L.A. & Cheu, J. (1978). Preparing for Tomorrow's World: An Alternative Curriculum Model for the Secondary Schools Paper. *First Annual Conference of the Education Section et the world Future*, Texas, 22nd October

Jiang, H. & Zhang, Q. (2014). Development and Validation of Team Creativity Measures: A Complex System Perspective. *Creativity and Innovation Management*, 23 (3), 264-275.

Johanson, G.A. & Brooks, G. P. (2010). Initial Scale Development: Sample Size for Pilot Studies. *Educational and Psychological Measurement*, 70(3), 394-400.

Johnson, S. & Johnson, R. (2009). *Conceptualising and interpreting reliability.* UK: Ofqual.

Kampylis, P. G., & Valtanen, J. (2010). Redefining creativity-analyzing definitions, collocations, and consequences. *The Journal of Creative Behavior*, 44(3), 191-214.

Kane, M.T. (2009).Validating the Interpretations and Uses of Test Scores. In Lissitz, R.W. (ed.) *The Concept of Validity Revisions, New Directions, and Applications* (pp. 39-64). United States: Information Age Publishing Inc.

Kaufman, J.C. (2006). Self-Reported Differences in Creativity by Ethnicity and Gender. *Applied Cognitive Psychology*, 20, 1065-1082.

Koretz, D. (2006). *Measuring Up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Krathwohl, D.R. (2002). A Revision of Bloom's Taxonomy: an overview. *Theory into Practice*, 41(4), 212-218.

Lambert, D. and Lines, D. (2000) *Understanding assessment: purposes, perceptions, practice*. London: Routledge Falmer

Lipman, M. (1987). Critical thinking: What can it be? *Analytic Teaching*, 8(1), 5-12.

Lipman, M. (2003). *Thinking in Education*. 2nd edn. Cambridge: Cambridge University Press

McPeck, J.E. (1981). *Critical Thinking and Education.* Oxford: Martin Robertson

McPeck, J. E. (1990). Critical Thinking and Subject Specificity: A Reply to Ennis. *Educational Researcher*, 19(4), 10-12.

Mednick, S.A. (1962). The associate basis of the creative process. *Psychological Review*, 69(3), 220-232.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741-749.

Moss, P.A (1994). Can there be Validity without Reliability? *Educational Researcher*, 23(2), 5-12.

Newton, P.E. (2012). Clarifying the Consensus Definition of Validity. *Measurement: Interdisciplinary Research & Perspective*, 10(1-2), 1-29.

Newton, D.P. (2014). *Thinking with Feeling: Fostering productive thought in the classroom*. New York: Routledge

Nisbett, R. E. & Wilson, T. D. (1977). The Halo Effect: Evidence for Unconscious Alteration of Judgements. *Journal of Personality and Social Psychology*, 35(4), 250-256.

Norris, S. P. & King, R. (1984). *The design of a Critical Thinking Test on Appraising Observations. Studies in Critical Thinking. Research Report No1*. Canada: Institute for Educational Research and Development.

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2017). Ha ha? Assessing individual differences in humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, 11(2), 231-241.

Plucker, J A. & Makel., M.C. (2010) Assessment of creativity. In Kaufam, J.C. & Sternberg, R.J. (ed.) *The Cambridge handbook of creativity* (pp.48-73). Cambridge: Cambridge University Press

Propp, V. (1968). *Morphology of the Folk Tale*. Translation by Laurence Scott. The American Folklore Society and Indiana University

Richards, J.C. (2005). *Communicative Language Teaching Today*. SEAMEO Regional Language Center

Rodari, G. (1996). *The Grammar of Fantasy: An Introduction to the Art of Inventing Stories.* Translation and introduction by Jack Zippes. New York: Teachers & Writers Collaborative

Rungo, M. A. & Jaeger, G.J. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, 24(1), 92-96. doi: 10.1080/10400419.2012.650092

Silvia, P.A. (2015). Intelligence and Creativity are Pretty Similar After All. *Educational Psychology Review.* 27 (4), 599-606. doi: 10.1007/s10648-015-9299-1

Sireci, S.G. (2009). Packing and Unpacking Sources of Validity Evidence. In Lissitz, R.W. (ed.) *The Concept of Validity Revisions, New Directions, and Applications* (pp. 19-37). United States: Information Age Publishing Inc.

Su, C.T. & Parham, L.D. (2002). Case Report- Generating a valid questionnaire translation for cross-cultural use. *American Journal of Occupational Therapy*, 56, 581-585.

Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2017). Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism. *International Journal of Science and Mathematics Education*, 15, 663-682.

Torrance, E. P., Ball, O. E. & Safter H.T. (2008). *Torrance Tests of Creative Thinking: Streamlined Scoring Guide for Figural Forms A and B.* Bensenville: Scholastic Testing Service Inc.

Treffinger, D.J., Young, G.C., Selby, E.C. & Shepardson, C. (2002). *Assessing Creativity: A guide for educators*. Florida: Center for Creative Learning

Weisberg, R. W. (2015). On the usefulness of "value" in the definition of creativity. *Creativity Research Journal*, 27(2), 111-124.

Yoon, C. H. (2017). A validation study of the Torrance Tests of Creative Thinking with a sample of Korean elementary school students. *Thinking Skills and Creativity*, 26, 38-50.

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Middle School Mathematics Teachers' Opinions on Feedback

Hacı Ömer BEYDOĞAN

To cite this article: Beydoğan, H.Ö. (2018). Middle School Mathematics Teachers' Opinions on Feedback, *International Journal of Assessment Tools in Education*, 5(1), 33-49. DOI: 10.21449/ijate.339410

Not: Türkçe versiyonu makalenin sonundadır.

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 33-49

http://www.ijate.net

e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Middle School Mathematics Teachers' Opinions on Feedback

Hacı Ömer BEYDOĞAN*

Ahi Evran University, Faculty of Education, Turkey

Abstract: During instruction, providing feedbacks improves students' academic achievements as well as motivates them to actively engage in lesson activities. Feedback is very important for teaching. Feedback is not only a functional tool to provide active involvement of the students to the learning process but also affects the academic success of the student. In this study, it is important to analyze in-service mathematics teachers' opinions on feedback. This study is conceptualized as a qualitative study. The data of this study included in-service teachers' responses to a semi-structured questionnaire, which created by the researchers. In-service teachers' responses to the interview questions were audio taped and later transcribed verbatim to conduct a content analysis. Twelve mathematics teachers working in 12 different schools in a central district of Kırşehir voluntarily participated in the study during the 2015-2016 academic year. The data of the study were obtained conducting face-to-face interviews with the mathematics teachers. Teachers' responses to the questionnaire items were analyzed thematically and classified under the following seven headings: style of the feedback, scope of the feedback, principles of providing the feedback, difficulties experienced when providing the feedback, reasons for providing insufficient feedback, the benefits of the feedback, and the significance of the feedback in learning. The results are presented in relation to the literature in the area. Teachers agree that it is not possible to complete students' gaps in Mathematics with the courses offered in the collective education system. Based on the findings some suggestions about the usage of feedback were provided.

ARTICLE HISTORY

Received: 21 March 2017

Revised: 28 July 2017

Accepted: 14 August 2017

KEYWORDS

Comprehensive feedback,

Feedback, learning process,

Mathematical conceptualization,

Mathematical operations,

1. INTRODUCTION

Feedback is an important external stimulus used by teachers to increase students' learning. Teacher-student interaction develops through the meanings that students attribute to external stimuli in teaching process. The feedback provided in the interaction process is significant when it meets the learning needs of students, when it is used for creating a suitable learning framework and when it is expressed as verbal and nonverbal stimuli that are appropriate for the developmental levels of students (Looney, 2005). Feedback is a stimulus that provides information which enables a student to focus on the problem area in a way that motivates his/her next action, that gives an opportunity for him/her to question whether he/she has understood, and that allows a student to evaluate both him/herself and his/her peers. Before giving feedback,

*Corresponding Author E-mail: hobeydogan@gmail.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.339410**

the teacher must identify the mistakes that student has made during the learning process and especially any misunderstandings, and must use observational and non-observational techniques effectively to do this. The type of assessment that is most effective in this situation is formative assessment (Kahl, 2005). Feedback that is effective in learning and teaching processes contributes positively towards education; however, it has negative impacts on a student's learning if it is misused or used carelessly (Hattie and Timperley, 2007). The feedback that takes place in all learning from the beginning of teaching period to the end forms the backbone of the formative assessment of student. Feedback can be used in a planned way by teachers in the teaching process, but sometimes it can also develop spontaneously. Planned feedback is the most important element of formative assessment (Black and William, 1998; Hattie and Timperley, 2007). From the student's perspective, feedback is a helpful stimulus that is used by student to verify, falsify what he/she did or to add to his/her knowledge and to provide information about his/her performance and understanding. In the learning process, students need such stimuli (Mory, 2004).

Teachers can use feedback in the teaching process through both open-ended and closed-ended questions, as well as providing instant explanations, facilitative actions and words to be learnt so that the student can complete his/her tasks. Feedback is also used for different purposes to influence learning. Teachers use feedback to fill the gaps between the level required of the student and the students' actual performance, to make students' learning more effective, to correct students' misunderstandings, procedural mistakes and erroneous strategies used during the learning-teaching process.

In formative assessment, feedback about the success of students plays an important role in integrating the learning and teaching process, in ensuring that students understand what they have learned, and in improving students' learning. The data obtained during the assessment process give clear guidance about the students' development and the steps and decisions that need to be made to progress to the next stage of the learning and teaching process.

Feedback not only has an effect on the academic achievement of student but is also effective in maintaining the active participation of students in the learning process (Brookhart, 2011) and in keeping them motivated (Wigfield, Klauda &Cambria, 2008). For this reason, it also influences students' competency in learning. The classical sense of feedback is communicating to students the knowledge of what corrections are necessary. However, in a contemporary sense, it consists of the information that student needs to have to know what to do next (Labuhn et al., 2010), the steps to be taken to improve his/her study skills incrementally and everything else to improve his/her work (Black & William, 1998; Hattie & Timperley, 2007; Sadler, 1998). Clark and Dwyer (1998), Foote (1999), Warden (2000) and Zimmerman and Martinez-Pons (1992) state that feedback is the most important source of information for students in correcting misconceptions, forming knowledge, supporting their metacognitive processes, improving their academic performance and increasing their motivation. According to Hattie and Timberley (2007), the main purpose of feedback is to emphasize the disjunction between students' current understanding and performance and the learning objective; moreover, it is one of the steps taken in order to encourage students to reduce this disagreement (Rakoczy, Harks, Klieme, Blum & Hochweber, 2013). In this context, feedback from teachers and students' peers is external guidance that explains how students can improve their performance (Butler & Winne, 1995), how to perform the tasks which they are required to, and how to monitor and evaluate students' progress (Stone, 2000). Feedback can be defined as statements or stimuli given to guide the student to the desired outcome.

When the research conducted in Turkey regarding teachers' feedback in the classroom is examined, it can be listed chronologically as follows: Yunt (1992) revealed that the use of feedback and correction together significantly increased overall marks. In the research

conducted by Saraçaloglu, Evin-Gencel and Çengel (2011), high school teachers' competencies during learning and teaching processes were examined from the teachers' and students' viewpoints, and teachers found their in-class behaviors adequate; however, students were found to have the opposite opinion. Şahin (2015) classified feedback as “explanatory”, “articulatory”, “diagnostic” and “remedial” in his research investigating the opinions of prospective teachers about the feedback applied in the learning and teaching process. In the study, it was determined that teachers used “confirmatory” feedback from time to time and the other types more frequently. In a study on the beliefs and behaviors of primary school mathematics teachers by Köğce and Baki (2012), it was stated that teachers generally used feedback in the learning-teaching process. However, they attributed different meanings to the concept of feedback. Köğce and Baki (2012) also stated that primary school mathematics teachers gave feedback according to students' personal characteristics as well as their performance and that some of the teachers exhibited some negative behaviors when they gave feedback. Eraz and Öksüz (2015) investigated the impact of the feedback given by primary school teachers to their students during extracurricular math activities on students' achievements and attitudes. In groups that had been given feedback, students' achievement and positive attitude scores increased significantly compared to the other group. Türkdoğan and Baki (2012) studied feedback techniques used by teachers by observing feedback about mistakes given by mathematics teachers at secondary school level. In the above studies, the effect of the teachers' feedback on the success and attitude of students, the types of feedback they used and the meanings they attributed to feedback were emphasized. This research aims to reveal the definitions, principles and approaches teachers use while they give feedback.

1.1. Feedback in Mathematics Teaching

To make mathematics topics that are perceived as abstract and difficult more easy, and to make them understandable for students, the functioning of teaching-learning process must be carefully observed. In general, it is crucial for teachers to interact with students in a sophisticated manner, to give appropriate feedback and to take steps to make sure students understand mathematical concepts and symbols from the beginning until the end of the teaching process. In this regard, it is necessary to find realistic solutions for mathematics teaching through monitoring the feedback teachers give and determining where it is deficient. In mathematics, it is necessary for every student to be able to think, consider and express his/her ideas using the mathematical symbols. For this reason, the every student’s achievement in mathematics is linearly related to their ability to read, understand and apply mathematical symbols.

In practice, every teacher is expected to show sensitivity in acquiring basic mathematical skills. Teachers must use feedback primarily to promote cognitive interaction with their students and guide them to the solutions to problems. Students need to get feedback to improve their skills with every step they take and every calculation they make during the teaching of mathematics. Such a process can only be achieved by maximizing the intensity of interaction with students. During this process, teachers try to increase students’ level of learning and preliminary knowledge, to develop appropriate teaching methods, to increase dialogue using different question types, and to encourage students to improve their competency to predict, analyze and interpret (Akyol, 2007). Feedback can thus be defined as the communication of information that is provided in order to improve students’ learning and to alter their thoughts and behaviors. Formative feedback includes not only the information given, but also, at the same time, the processes and activities that will support the learning of the student.

In teaching mathematics, how students react, in terms of determining and correcting the misconceptions formed during the acquisition of mathematical concepts, mistakes in process steps, the completion of the process and the interpretation of processes are all of great

importance. Mathematics teachers try to perfect the process through the feedback they use in teaching-learning process. It is not possible for a student to learn mathematics in a process that does not include feedback, because feedback forms the basis of mathematical thinking, conceptualization and correction.

According to Şantagata (2002), teachers' feedback in teaching mathematics occurs through correcting a student, giving hints/clues, repeating the question, asking for reasons, giving hints to a different student, asking a question indirectly, choosing the correct answer, requesting the correct answer from students, and finding the correct answer using the students' attempts to answer the question. In the teaching process, teachers can help students gain appropriate skills by determining the mistakes that have been made and by finding the correct answer through these mistakes (crosschecking). In a classroom, students can gain meaningful learning and skills through their mistakes (Nordstrom, Wendland & Williams, 1989). Çevikbaş and Argün (2016) have determined that the types of feedback given by mathematics teachers to wrong answers had both positive and negative effects on self-esteem.

In studies conducted abroad, Roschelle et al.(2010) investigated the effects of technology-supported feedback on the mathematics learning of students in groups. While Labuhn, Zimmerman & Hasselhorn (2010) focused on the impact of feedback on perceptions of self-efficacy and the problem-solving performance of students, Naroth (2010) tried to determine the effect of teacher's structured feedback on improving students' mathematics learning. Carvalhoa, Santosa, Conboya and Martinsa (2014) emphasized the role of teacher feedback in eliciting perceptual differences among students in their research conducted with 179 students. Duhon, House, Hastings, Poncy and Solomon (2015) investigated the contribution of feedback to mathematics learning in terms of timing and explanatory features.

Feedback in the teaching-learning process can serve to help students understand mathematics, to read mathematical symbols and to correlate processes correctly, if it meets students' needs. The effective use of the process of conceptualization, which forms the basis of mathematics teaching, is something that directly affects the learning of the student. Teachers usually give feedback during this process. The correct use and correlation of mathematical symbols form the basis for students' understanding of mathematical content and their cognitive development.

As feedback has such an important function in students becoming mathematically competent, this study aimed to determine how teachers perceive feedback, to investigate their thoughts about feedback and behaviors when giving feedback thematically, and to evaluate these in light of the current literature.

For this purpose, the question "What are the opinions of mathematics teachers about the feedback they give in class?" was posed. An answer was sought to the following questions in accordance with this problem statement.

1- In your opinion, what is the function of feedback in the teaching of mathematics?
2- What kind of approach do you follow when you give feedback during the teaching process?
3- What kind of feedback do you generally give to students during the learning and teaching process, and for which content?
4- Which basic principles do you use when you need to give feedback to students?
5- What are the difficulties in giving students feedback about mathematics?
6- What are the benefits for mathematics teachers of the feedback given in the teaching process?

2. METHOD

The qualitative research method was used in this study. Qualitative research can be defined as a research process in which qualitative data collection methods such as observation, interview and document analysis are used and a qualitative process is followed in order to reveal facts and events in a natural and realistic way in a natural environment (Yıldırım & Şimşek, 2008). The data of the study were obtained from voluntary face-to-face interviews with 12 mathematics teachers who had been working in mathematics teaching for five years or more at secondary school level.

In the interview, a recording device was used to prevent data loss. Participants were informed that a device would be used. It was stated that participants could listen to the recordings at the end of the interviews, and, if necessary, the opinions expressed in the recording could be removed partially or completely if requested. Thus any potential problems participants may have had about being recorded were eliminated. Throughout the research, participants were provided with an environment which made them feel comfortable and at ease and thus able to express their views honestly. During the interview, participants were asked not to be influenced by the researcher while answering the questions. In order to increase the reliability of the research, teachers were asked to specify their role in the class. Individuals who were data sources were clearly defined, and the social environments and processes formed in the research process were also defined.

In the interview, six semi-structured questions developed by the researcher were addressed to the participants and responses were recorded. Participants were asked to answer again some questions asked in previous face-to-face interviews to check if they gave the same answers. This was to try to ensure consistency in the information collected. Consistent statements were included in the analysis. Interviews with the participants covered a period of four weeks. Verbal explanations recorded in the interview were written down, assessed and analyzed. The content obtained from the interview was thematically analyzed.

The study group consisted of 12 mathematics teachers with more than five years of teaching experience working in 12 different schools in the central province of Turkey during the 2015-2016 academic year and they participated in interviews voluntarily. Data obtained in the interviews were recorded and then analyzed and written down. Results obtained were correlated with data from literature, discussed and presented.

2.1. Analysis process

In the analysis of the opinions of the mathematics teachers, groupings were made according to the similarity of the statements, teachers who had been consulted were given a code number (e.g. K1, K2...) and explanations were given. Similar items in statements were grouped together thematically and “themes” were named appropriately. Concepts that constituted themes were grouped among themselves to ensure consistency, the themes were evaluated for consistency alongside other themes, and tested to see whether they formed a coherent whole. The suitability of the findings was compared with previous studies. Themes were explained and interpreted by the deductive or inductive methods according to the situation. The findings were reviewed by the participants and found to be realistic.

The consistency of the research findings with predictions made was taken as a basis. In order to obtain external validity in the data obtained, details of the period of investigation from the preparation of the data collection tool to the application and analysis phase were explained. The attempt was made to determine the consistency of the findings with the practical realities by comparing findings with the literature. The aim was to describe explanations in detail in order to be able to test the research against other research. The participants (mathematics teachers) were interviewed again and the findings were confirmed by being shared.

3. FINDINGS

A number of themes were revealed in by the findings obtained in the research. Themes were grouped under topics such as the “place of feedback in teaching”, “type of feedback”, the “content of feedback”, the “principles of giving feedback”, “difficulties in giving feedback” and the “benefits of feedback”. *The following points stand out under the theme “feedback's place in teaching”*

Eliminating uncertainty	K12
Combining old and new knowledge	K11
Providing enduring knowledge	K1, K11
Reducing error	K4
Facilitating understanding	K7, K12

Feedback is external stimuli that teachers use to increase learning in their students during the teaching process. Students interact and understand the appropriate meaning from these external stimuli. Feedback provided during interaction is important if it meets the learning needs of student, is used to make the learning frame appropriate and is expressed through verbal and nonverbal stimuli that are appropriate for the level of development of the students (Looney, 2005). According to its function, feedback is stimuli that provides students with necessary information, that allows the student to take action in the next required task, gives him/her the opportunity to question whether he/she understands it or not, and allows him/her to evaluate him/herself and his/her peers.

Teachers plan the learning process with their students and create teaching objectives. They recognize the deficiencies and gaps that arise during students’ learning and give feedback in a way that will remove these gaps. Teachers use different feedback strategies to fill the gaps in students’ learning. Mathematics teachers were asked the open-ended question, "What, in your opinion, is the function of feedback in mathematics teaching?" The responses of the teachers are given below:

K12- Feedback is given to students to make them understand mathematics topics better. It is a fact that a student asks the teacher a question and gets an answer whenever there is a situation that confuses the student. Feedback helps to answer the student's questions. When it is not given, a student cannot understand the topic fully, and when the topic is talked about a week later, the student actually doesn’t have any idea anymore what it’s about. As time goes on, it becomes increasingly difficult for the teacher to help that student.

K11- We give feedback especially after an exam to make a child combine old knowledge with new knowledge. It is actually indispensable to give feedback. It is important to see whether the behaviors they have learned are permanent or not and whether or not they’ve gained any new behaviors.

K1 - When a student does something wrong or something right, they’ll always remember it. If they can remember something, those things are more valuable to them.

K4 - When we give feedback that illuminates the points that a student doesn’t understand, the student makes use of them. The child doesn’t make the same mistake again.

K7 - We give as much explanation as necessary to prevent students from misunderstanding mathematics topics, misinterpreting them, and getting the wrong answer. However, if the student does not have an aptitude for numeracy, we can’t get the results we want.

Teachers stated that feedback provided during the teaching-learning process contributes to a better understanding of mathematical content, integrates the students’ prior knowledge with new knowledge, removes any uncertainties from students’ minds, increases the durability of

learning, and contributes towards providing the knowledge and skills students need. The following points stand out under the theme of “type of feedback”

Concentrating on the result	K5
Demonstrating mistakes	K5, K8
Benefitting from analogy	K8
Using I language	K8, K2
Asking questions together	K11, K5, K7
One-to-one attention	K12, K1
Reviewing the process and the result	K12, K2, K5, K1

When messages are shared in an effective communication process, it is important for both the student and the teacher to understand the content of the message correctly. The clarity of the message is closely related to how feedback is given. The answers teachers gave to the question "What kind of approach do you follow when you give feedback during the teaching process?" are as follows:

K5 - When I give an answer to student, I do not say "You did it wrong", I say "If you solve the problem in this way, you get the wrong answer, but if you solve the problem in that way, you can get the right answer.” Thus I give results-oriented feedback.

K8 - Instead of saying "This is not true, how could you do this?” I say "This part is correct, but the rest is wrong." I get the students to discuss it among themselves to find the right answer through the wrong answer. It is good practice for students who make similar mistakes in the class. Students don’t enjoy making mistakes, but do like correcting the mistakes and discussing them in the classroom.

K11- In mathematics, I prefer to give feedback to children individually, because the point at which each student has a problem is different. Every student brings me his/her solution one by one, and when I see the point where he/she has a problem I say, "Look, you need to look again at that part of the solution." This is more useful for child because I’m explaining his/her mistake.

K12- The student shouldn’t worry. For example, even if I have to tell some students three times, and they say, "I can’t do this anymore", I push them and say, "Good job, try harder", which reduces the pressure that they will fail. At other times, I say, "You see that you really have to do this in this way, don’t you?" After that, child's attention on the course increases. In short, the environment in which the feedback is given and the type of feedback given are very important.

K5 - When giving feedback, you can humiliate the child or you can give him/her advice gently. I think that if I give gentle advice the result will be better. You know the saying, "Kindness opens every door.” I make the student focus on solving the problem by asking simple questions. I try to meet the learning needs of students by providing hints to remind them of the rules when it is necessary.

K7 - I assess the situation first, and when the answer the child has found overlaps with the solution I ask, "How did you get this answer?" and I want the children to explain the situation. If the student really did the right thing, I thank him/her in front of the other students in the class. If he/she made a mistake, I make him/her solve the problem by questioning it together.

K2 - Children solve problems in different ways. For example, I've just experienced it today, the child had written the workings down differently, but the answer was the same, so there was no difference. They may ask, "Teacher, can you look at this? I found the answer in a different way?" I go and check it immediately, I respond to them instantly, they like this, and their interest and engagement increase. In short, children want special attention.

K1 - During the exercise, I go to students and give feedback constantly, because they feel that I am paying them attention when I go and check them.

K2 - I use body language. I don't shout a lot. For example, when I make this hand gesture (when I cross my hands over my chest), they all lean back. I don't get angry and I give them the explanation.

K5 - In the classroom, the 'U' table layout is very important. That is, children have to sit in a U-shape, they have to see each other. If students in all classes sat in this layout, they would monitor each other's work, interact and learn from each other.

K5 - Students in the class sometimes don't have many questions about how to solve a problem so I guide them in their work. However, I can get feedback from the students when I make them solve questions on the board.

K1 - The examination system should be completely changed. Students shouldn't be prepared for exams, they should learn the basic knowledge and skills, and activities for this should be increased. Unfortunately, exam anxiety and pressure turn a mathematics teacher into someone who only solves problems!..

Teachers agree that feedback given to students needs to focus on their learning needs. They are aware of the fact that the student likes and enjoys this interaction and makes use of it, if the student is shown where their answers are lacking, inaccurate and inadequate. They emphasize the importance of organizing students' seating to ensure that feedback is given directly and clearly and that multiple interactions are provided for. Teachers think that in math courses, instead of the teacher being the only the person who is active and solves a problem, he/she will succeed better in teaching mathematics by basing it on student-centered activities. The following points come into prominence under the theme of feedback's content. The following points stand out under the theme of "content of feedback"

Explaining, demonstrating and making students solve the problem	K7, K5
Making students think	K2
Using similarities and differences	K5
Highlighting details	K11
Devotion	K3

Teachers' answers to the question "What kind of feedback do you generally give to students in the learning and teaching process, and for which content?" are as follows:

K7 - I explain, demonstrate, solve the problem and give enough examples about the topic for the student not to have any questions in his/her mind. If necessary, I solve extra problems or let the student solve it him/herself. It is very difficult for children to learn math if we do not do this.

K2 - If you make a child memorize mathematics, the child can't do anything. Mathematics can't be memorized, the child needs to think a little abstractly. It is very important to associate mathematical concepts with the correct meaning and appropriate symbols and internalize them.

K5 - I think that it is necessary that a student not be confused in order for them to make the correct inferences and find the correct solution. Incorrect processes and inferences result in missing information or misconceptions in the child's mind. I show them how to solve a problem in different ways if possible, and if necessary, I check it. I want students to try similar ways. A student whose mind is not confused does not make mistakes, and does not make wrong inferences.

K3 - How much information can you give to a student when his/her level is very low, that is, when the basic skills are lacking? However much you explain the topic, he/she does not understand because the child is not able to understand. At the same time, we don't really

focus on this, because if we focus on that child, maybe half an hour or an hour passes, and if we spend time with that child, we ignore the other children.

K11- This year, I'm teaching 5th grade math classes. In lessons, I give the solutions to all exercises one by one and answer the children's questions. I don't have any time management problems because I have 5 hours of classes, 2 hours of an elective course, and 4 hours of other courses. I have the opportunity to answer each student's questions during this time.

Mathematical subjects require abstract, reasoning and deduction due to their content. It is clearly stated that this course cannot be taught through rote learning, and that students need feedback about their activities to make up anything lacking. It is stated that teaching mathematics involves a gradual process, progressing from concepts to processes. The teachers emphasize that it is necessary to give comprehensive feedback in order to make up conceptual and procedural deficiencies in the students' knowledge. The following points stand out under the theme "principles of behavior in giving feedback"

Bringing existing competencies to the forefront	K8, K3
Starting with simple examples	K4
Using I language	K1, K11
Demonstrating mistakes	K1, K7, K11
Using existing correct answers to motivate	K3
Communicating individually	K12
Time management	K5

Teachers' answers to the question "Which basic principles do you use when you need to give feedback to students?" are as follows:

K8 - If a part of student's answer is correct, bringing it to the forefront and saying "Well done, look! You did this part correctly but the rest will not be like this, it will be like that" increases the self-confidence of child.

K11- I try to make the statements I make clear, and to demonstrate the answer to child calmly and slowly for him/her to understand. I give simpler examples to the students with lower levels and try to explain the example I give a little slower.

K1 - Saying "Why did you do it wrong? It is not like this, that way ..." sternly reduces success of the student gradually.

K7 - When a student makes a mistake, I call him/her to the board to notice his/her mistakes, and say "Look when you do this in this way, you make a mistake, but in the way you can find the correct result". After student finds the right solution under my control, I say "Solve another one on your own.".

K3 - If student gives correct answer in some parts of the question, saying "These parts are correct", child's self-confidence increases and it is beneficial. That is why, I tell him/her correct answers.

K12- In correcting wrong behaviors or consolidating correct behaviors, the teacher must first know the student with the student's name, surname, personality, something that he/she can do and cannot do. If the teacher really knows his/her student, he/she must always make student feel valuable. This improves the success, otherwise feedbacks are meaningless. It is not enough for you to tell what you know. You should win that student's heart. When you do not know the student, you cannot help him/her saying "Hey boy or girl…". But if you know his/her family, if you have family's phone number, in such a case you can make family to get into situation immediately and you will be in dialogue with the family. So that student says "I have no place to escape".

K5 - Timing should be very appropriate when giving feedback.

K4 - First, we need to measure a student's foreknowledge about the topic, in fact. If the student does not have that preliminary information, you should give it to the child as a teacher.

Through the research studies carried out, it is necessary to reach a number of general truths and conclusions from the knowledge, experience and observations obtained. When teachers' opinions are examined in order to determine how well they are committed to the principles when they give feedback, the emerging opinions are as follows: Teachers pay attention to the need to know their students with different techniques. In order to encourage the student, first they can talk about what they can do and their competencies, then talking about the lack of knowledge will empower the student balancing him/her emotionally and build self-confidence. In doing so, they have a common view that a soft style involving "I language" should be adopted instead of using hard, accusative language. It is emphasized that students should be prepared to learning, that feedbacks should be chosen according to the level of student and should be presented with concrete examples and it should be committed to the general learning principles. The following points come into prominence under the theme of difficulties in feedback

Lack of foreknowledge	K11, K1
Not lowering him/herself to the level of students	K11, K1, K8
Not sparing time for student	K11, K10
Having difficulties in perception, understanding and solution	K9, K1
Hyperactivity and resisting to learning	K1, K8
Content intensity	K12

In teaching-learning process, teachers' opinions about "difficulties in giving feedback to student about mathematics" are as follows:

K5 - I have difficulty in responding to students when there is a lack of foreknowledge about the topic.

K9 - Feedback depends on student's level, students who have a good level receive feedback, especially students in middle level respond and receive feedback. It is very difficult to establish that interaction with low-level students.

K10 - It is very difficult to give feedback to every student in the class because the number of students in our classes is very high and the interest in some of our students is very low. Each child must first perceive the question and then answer. We can help him/her at some point that he/she cannot solve. But the child already does not want to solve the question and he/she also resists. It is not possible to give feedback at this point. If you give something to child, you can reach a result. So you interact mutually. There is no result obtained from one-way interaction.

K1 - Students are active, constantly hustling inside and out, and bickering with each. I mean it doesn't zone out for them whether they are talking to each other, making an explanation about a problem, or making friendship with each other. They convert the act of learning something with a feedback given to a group into an act of resistance. They utilize the opportunity of saying "look, he couldn't do it" or "he couldn't solve the problem again" whenever they have a chance.

K11 - If the basis of education of the students is weak, they cannot understand the topic even if I explain too many times and simplify it. For these students it is necessary to go back 3 or 4 years. But unfortunately it is not that easy; it is really difficult. For example, some students from the 5th grade are actually at the 1st or 2nd grade. I am aware that they also have difficulty ... I am trying to overcome this problem with them by asking a successful student to explain a topic to one of his friends. This duty is beneficial for the successful student, because he grabs a chance to consolidate his knowledge. It is also beneficial for the other student, because I don't have a chance to spare enough time for each student. I try to

overcome this problem by giving such a duty to an older sister or brother, or by talking with parents.

K12 - I have a difficulty in giving enough feedback to students. Because there are so many topics they have to learn, and I have to finish the curriculum. I have to explain the topics one more time before the pilot exams are done. We work in an environment of a big competition. The school is one of the most successful schools in the city. The administrator knows that my class has a success rate of 94%. If the success rate of my class drops from 94% to 92%, he says that success of my class has fallen and that I am responsible for it. Instead of giving detailed feedback to each student, I try to save both time and success by giving feedback processionally. At the same time, I show my students some of their behaviors. They bear fruit by realizing their mistakes and seeing the correct behaviors from their teacher.

K8 - For the relationship between the sender and the receiver to be healthy, you have to know when and how much message the receiver can receive. Learning new information is not possible for students without learning the basic information they need to learn in subclasses. Because of this, I sometimes lower the level of feedbacks. At this time, while acting so, other students in the class make fun of that student, or a communication gap occurs in that class. For example, I ask a simple, basic knowledge to one side of the classroom, but children who have basic knowledge on the other side react to children who do not. They say "Come on, don't you know this?", and as a result of it, the students who cannot understand the topic turn in on themselves. And thus, it becomes more difficult to receive feedback from those students.

K1 - It is not possible to be able to focus on the student with low level too much. We can pay attention in our leisure time or break time.

Mathematics teachers are found to have difficulties when giving feedback because of weakness of students' readiness on mathematics, lack of foreknowledge, low levels of learning, fear of failure and lack of self-confidence due to content intensity and abstractness of mathematics topics, high number of students in classes and not allow enough time for each student. The following points come into prominence under the theme of benefits of feedback.

Recognition	K9, K2, K7, K5
Increase in attention	K9
Positive attitude	K8, K7
Self-confidence	K9
Self-control	K5

Answers of teachers to the question of “What are the benefits of feedback given in teaching process for mathematics teachers?" are as follows:

K9 - It is important to give feedback to the students, because when the student individually understands the correctness or wrongness of the result, the student notices that, the teacher pays attention to me. The interest of the student increases.

K2 - It is important to give individual feedback to the students. When I see something that the child asks, I say "Look at it carefully". It is more useful for that child to talk directly to his/her mistake.

K8 - Feedback is very useful in terms of participation of students in other courses, self-confidence, approach to the course and teacher. We must give feedback to the student making student notice his/her mistakes and rights.

K5 - When we give feedback to the students, we control ourselves first of all. So, knowing what we give in courses, how much we give, how much student understands, which points are not understood is a chance to focus on the topic again.

K7 - Feedback helps the student understand the topic completely. It removes the remaining question marks from students' minds.

K2 - Where did he/she make a mistake? What is the source of this error? What if he/she does not know the addition, subtraction, multiplication and division processes in mathematics? Does not he/she know the formula? Did he/she ever understand? These are very important to us. Because the feedback is a key for us.

Teachers think that it is more correct to see how much the student understands the topic explained, to repeat the things they do not understand, and to tell the students' deficiencies directly to the face of students. Teachers agree that feedback is useful both in their own evaluations as teachers, forming self-confidence in students, making students participate in teaching, and ensuring students to focus on learning.

4. DISCUSSION AND CONCLUSION

In the interviews with mathematics teachers, the structure of the feedback they gave in the class was studied and investigated thematically. Themes were grouped under topics: the place of feedback in teaching, the type of feedback given, the content of the feedback, principles of behavior when giving feedback, difficulties in giving feedback and the benefits of feedback.

With reference to the opinions of teachers participating in the study, the place of feedback in teaching was evaluated under the headings “removing uncertainty in learning”, “combining old knowledge with new”, “providing enduring knowledge”, “reducing errors” and “facilitating understanding”. In terms of the place of feedback in teaching, the majority of teachers were consciously concerned about not giving comprehensive and appropriate feedback to their students. They tried to explain these concerns with reasons such as having large classes and the low level of basic knowledge of the students. In mathematics teaching, a teacher gives feedback by correcting mistakes, giving hints, repeating the question, asking for explanations, giving more hints to different students, asking the question indirectly again, requesting the correct answer from students and finding the correct answer using the students’ attempts (Şantagata, 2002). In some cases, teachers also help students gain meaningful learning and skills in the classroom by using their students' mistakes (Nordstrom, Wendland & Williams, 1989). Köğce and Baki (2012) found that primary school mathematics teachers give feedback according to the students' personality as well as performances. In addition, Eraz and Öksüz (2015) found in their research that scores for student achievement and positive attitude were significantly higher in the groups that received feedback.

The mathematics teachers here stated that they gave feedback that was adequate for each student's needs. Şahin (2015) revealed that teachers frequently use feedback to attract a student's attention, to motivate them, to inform them about the goal, to give hints, and to encourage required behaviors.

In terms of the theme of the “type of feedback given”, the mathematics teachers' statements were about making students concentrate on the result, demonstrating mistakes, using analogy, using I language, asking questions together, one-to-one attention and guiding the process and the result.

It seems that the teachers here made an effort to determine why they gave feedback and what the problems were. It is very useful for teachers to ask simple questions to find out if students know the topic, to remind them about the rules for solving the problem, to give hints about the solution and to meet their learning needs. When a teacher sees a conclusion that is in conflict with how the problem should be solved, he/she can ask how this situation emerged and can ask student to explain the situation. In this way, students can gain the ability to think and experience things deeply and gain the ability to re-focus. Mathematical thinking is a process that requires intense cognitive activity. In the process of cognitive learning, feedback is necessary to remove any contradiction in students’ thinking. It is necessary for the teacher to use more than one method and technique to solve problems and for them to get students to use

different techniques to find a solution, so they can reflect on this and learn. The results of this research correspond to conclusions of Türkdoğan and Baki (2012) about feedback given by teachers about mistakes. The feedback that teachers give in this way is significant.

With regard to the theme of the "content of feedback", teachers focused on the necessities of explaining, demonstrating and making students solve the problem, thinking about the problem, using similarities and contrasts in problem-solving, emphasizing details and sparing time for each student. It can be understood from the answers the teachers gave that they are conceptually confused about the content of feedback. Some teachers perceive feedback as answering the question, some of them perceive it as a guide to facilitate the learning of student, or only in terms of body language.

Reminding students of the mathematical concepts and procedures that the student has learned before when he/she makes a mistake, and giving them stimuli which enable them to make new inferences can be seen as an effective way of giving feedback. Essentially, making students feel or recognize that they have made a mistake in some topics is the most effective way to give feedback, because intuitive-based learning has the property of encouraging intrinsic motivation and effort.

Associating concepts that have similar features, linking processes and problems with topics previously learned in order to make inferences about results may be effective feedback techniques for some students. This process must be supported by what the student does. In this case, the teacher may help the student to recreate the problem in a way that enables the student to internalize the process. If the student indicates that he/she understands the topic, but continues to do the wrong thing, the teacher can try to find a solution by making gradual transitions from solving a simple problem to solving a complex problem by simplifying the question. In solving complex problems, the teacher can try to teach by dividing the problem into smaller, but still meaningful parts. The main aim is that the student can answer the question on his/her own. When stimuli are given to encourage students to progress correctly at their own pace, students who continue to make mistakes can be provided with solutions to similar questions that are easier to understand, and it can be ensured that the student understands the logic of the process and can make meaningful associations.

With regard to the theme of "principles of behavior in feedback", teachers' opinions were focused on bringing existing competencies into the forefront, starting with simple examples, using I language, demonstrating mistakes, using existing correct answers to motivate, communicating individually and time management. Every reaction given to students must be time-limited but consistent. This limit and consistency need to take students' individual differences into account. In a collective education system, extra explanatory feedback should not be given to only one student, while others do not receive any, or receive more limited feedback.

In mathematics teaching, the teacher is expected to start the topic from the place agreed, to continue teaching, to follow strategies based on integrating what students understand and what they will learn. It is necessary for the teacher to wait for the student to explain the answer if the student knows the answer but also thinks that he/she has done it wrong, and for the teacher to get the student to repeat how they have worked it out to see if this is the case. Mathematics teaching requires patience and the management of students' thinking. However, in some cases, it can turn into a major problem for the teacher if they do not have enough time to give feedback, if teaching is unplanned, if the content is too intensive, and if the students are not ready.

With regard to the theme of "difficulties in giving feedback", teachers focused on lack of previous knowledge, not being able to match the level of students, not having enough time for students, having difficulties in perceiving, understanding and solving problems, students' hyperactivity and resistance to learning and the intensity of the content. In the context of the

difficulties experienced in feedback, it is revealed that serious problems are being experienced, which originate from teachers, students, the topics and the system of teaching.

The fact that mathematics is abstract and the fact that the topics taught are not as concrete as others demand that the teachers give clear, understandable, comprehensive and explanatory feedback. Since mathematics is a field entirely based on symbols, which require a conceptual model and processes closely related to this, it is necessary to try to make the mathematical concepts that are the basis of these symbols meaningful when teachers first begin teaching topics.

The findings obtained in this study show that the teachers are both highly sensitive and also very positive about how feedback functions. However, the student's level of readiness, past experiences, interaction with classmates and being in a competitive environment instead of a solitary environment seem to be factors that increase difficulties in motivating students. This makes it difficult for the teacher to provide adequate and comprehensive feedback, to give feedback on the basis of the students' performance, and to give feedback in an ordered way. The fact that the classes are overcrowded, the lack of a physical setting or equipment suitable for multiple interactions, the intensity of mathematics content, the fact that this content sometimes does not fit into the curriculum and the lack of time for the mathematics course all limit the possibility of giving comprehensive feedback to students.

It is difficult to identify the learning deficiencies that occur in the students in environments where the teacher is the guide but the students are not able express themselves. Giving oral explanations to students who lack conceptual knowledge and reminding them of rules do not by themselves make feedback effective. For students with inadequate knowledge of procedures, explanations about the correct use of symbols or where and how the concept is used in everyday life are not enough to fill in what they lack. Teachers should give the feedback in a manner that complements the teaching process according to the type of information that the students need, the process, the time available, and the concrete realities of daily life. A mathematics teacher who wants to teach successfully needs to know his/her students' level of readiness, motivation, personal characteristics, expectations, habits and attitudes before giving feedback. It is not possible to wait for a student who is not aware of the knowledge they lack to acquire it. Feedback given by teachers in the teaching-learning process should serve to give students awareness of their own deficiencies, their mistakes and their own areas for improvement. If feedback is comprehensive, specific and principles-based and increases interaction, it can contribute to this during the teaching process. Within this context, mathematics teachers

- should take a multi-faceted approach to the characteristics of their students and try to give appropriate feedback according to their individual differences.
- should give corrective feedback using the methods that enrich the thinking process such as thesis-antithesis, going from rule to example and deduction-induction, consolidate the correct actions of the students and correct mistakes by using them as an opportunity for improvement.
- should first give concrete examples to make students understand and internalize those mathematical concepts that constitute the basis for each new topic taught to students.
- should closely follow the learning process of their students and determine what kind of deficiencies they have in what areas and give complementary feedback.
- should use reinforcing, supporting and directing feedback in learning environments where students can participate more effectively and express their ideas better.

It is easier for teachers to identify what a student has in his/her mind and give appropriate feedback in a learning environment where the student is active and explains his/her ideas.

Teachers should also try to use feedback at the beginning of their teaching to determine what misconceptions their students have about mathematics, to amend any deficiencies and to eradicate their students' mistakes after the solution and answers to a question have been examined.

Acknowledgements

This work was supported by Ahi Evran University, BAP (Project No. EGT. A3.16.15)

5. REFERENCES

Akyol, H. (2007). *Vygotsky, Piaget ve Yapılandırmacı Okuma Eğitimi*. VI. Ulusal Sınıf Öğretmenliği Kongresi Bildiri Kitabı, Eskişehir.

Black, P. & William, D. (1998). Assessment and classroom learning. *Assessment Education, 5 (1),* pp.7-74.

Brookhart, S.M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 3*(1), pp. 3-12.

Clark, K., & Dwyer, F. M. (1998). Effect of different types of computer-assisted feedbacks strategies on achievement and response confidence. *International Journal of Instructional Media, 25*(1), pp.55-63.

Carvalhoa, C., Santosa J., Conboya, J. & Martinsa D. (2014). Teachers´ feedback: Exploring differences in students´ perceptions. *Procedia - Social and Behavioral Sciences,* 159, pp. 169-173.

Crooks, T.J. (1988). The impact of classroom evaluation on students. *Review of Educational Research,* 5, pp.438-481.

Dempsey J.V., Litchfield B.C., & Driscoll M.P., (1993). Feedback, Retention, Discrimination Error, and Feedback Study Time, Journal of Research on Computing in Education, 25: 3, pp. 303-326.

Duhon, G., House, S., Hastings, K., Poncy, B., & Solomon, B. (2015). Adding immediate feedback to explicit timing: An option for enhancing treatment ıntensity to ımprove mathematics fluency. *Journal of Behavioral Education, 24*(1), pp. 74-87.

Eraz, G., & Öksüz, C. (2015). Sınıf öğretmenlerinin öğrencilerin ders dışı matematik etkinliklerine ilişkin uyguladıkları geribildirimlerin etkisi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 36,* ss.105-119.

Erişen, Y. (1997). Öğretim elemanlarının dönüt ve düzeltme davranışlarını yerine getirme dereceleri. *Kuram ve Uygulamada Egitim Yönetimi Dergisi, 3*(1), ss. 45-62.

Foote, C.J. (1999). Attribution feedback in the elementary classroom. *Journal of Research in Childhood Education*, 13(2), 155-166.

Hattie, J. & Timperley, H. (2007). The power of feedback, *Review of Educational Research*, 77 (1), pp. 81-112.

Kluger, A.N. & Denisi, A. (1996). The Effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 2(2), pp. 254-284.

Köğce, D. & Baki, A. (2012). İlköğretim matematik öğretmenlerinin geribildirim kavramına ilişkin inanışları, X. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi, 27-30 Haziran, Niğde.

Labuhn, A.S., Zimmerman, B.J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning, 5*(2), pp. 173-194.

Mory, E.H. (2004). *Feedback research revisited*. In D. Jonassen, (Ed.), Handbook of Research on Education Communications and Technology (pp. 745-783). Mahwah, NJ: Lawrence Erlbaum Associates.

Manouchehri, A. (2007). Inquiry-discourse mathematics instruction. *Mathematics Teacher, 101* (4), pp.290–300.

Manouchehri, A. & St. John, D. (2006). From classroom discussions to group discourse. *Mathematics Teacher, 99* (8), pp. 544–551.

Naroth, C. (2010). Constructive teacher feedback for enhancing learner performance in mathematics.[serial online]. n.d.; Available from: Networked Digital Library of Theses & Dissertations, Ipswich, MA. 21 Ağustos 2016 da ulaşılmıştır.

Nordstrom, C.R., Wendland, D. & Williams, K.B. (1998). "To err is human": An examination of the effectiveness of error management training, *Journal of Business and Pschology*, 12, 3, pp. 269-282.

Kahl, S. (2005). Where in the world are formative tests? Right under your nose! *Education Week*, *25* (4), 38.

Looney, J. (Ed.). (2005). *Formative assessment: Improving learning in secondary classrooms*. Paris, France: Organisation for Economic Cooperation and Development.

Peker, R. (1992). Geri bildirimin üniversite öğrencilerinin ölçme ve değerlendirme dersindeki başarısına etkisi. *UludağÜniversitesi Eğitim Fakültesi Dergisi*, 7(1), ss. 31-39.

Rakoczy, K., Klieme, E., Bürgermeister, A. & Harks, B. (2008). The interplay between student evaluation and instruction. grading and feedback in mathematics classrooms. *Zeitschrift für Psychologie, 216*, pp. 110-123.

Rakoczy, K., Harks, B., Klieme, E., Blum, W. & Hochweber, J. (2013). Written feedback in mathematics: Mediated by students´ perception, moderated by goal orientation. *Learning and Instruction*, 27, pp. 63-73.

Roschelle, J., Rafanan, K., Bhanot, R., Estrella, G., Penuel, B., Nussbaum, M., & Claro, S. (2010). Scaffolding group explanation and feedback with handheld technology: Impact on students' mathematics learning. *Educational Technology Research and Development*, (4), pp.399-404.

Sadler, D.R. (1998). Formative assessment: revisiting the territory. *Assessment in Education, 5*(1), pp. 77-84.

Saracaloğlu, A.S., Gencel, İ.E. & Çengel, M. (2011). Öğrenci ve öğretmen görüşleri açısından lise öğretmenlerinin öğretme sürecindeki yeterlikleri, *Adnan Menderes Üniversitesi Eğitim Fakültesi Eğitim Bilimleri Dergisi, Aralık, 2 (2), ss.77-99.*

Stevenson, C.E., Heiser, W. J. & Resing, W. C. M. (2013). Working memory as a moderator of training and transfer of analogical reasoning in children. *Contemporary Educational Psychology*, 38(3), pp.159-169.

Stone, N.J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*, pp. 437-475.

Şahin, M. (2015). Öğrenme ve öğretme sürecinde uygulanan dönüt etkinliği ile ilgili öğretmen adaylarının görüşlerinin incelenmesi, *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 15(1), ss.*247-264.

Santagata, R. (2002). *When student make mistake: Socialization practices in Italy and the United States*, Doctoral Dissertation, Los Angeles: University of California, Philosophy in Psychology.

Turkdoğan, A. Baki, A. (2012). İlköğretim ikinci kademe matematik öğretmenlerinin yanlışlara dönüt vermede kullandıkları dönüt teknikleri, *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *45*, (2), ss.157-182.

Warden, C.A. (2000). EFL business writing behaviors in differing feedback environments. *Language Learning*, 50 (4), pp .573–616.

Wigfield, A., Klauda, S.L., & Cambria, J. (2008). Influences on the development of academic self-regulatory processes. In B.J. Zimmerman, & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp.33-48). New York: Routledge.

Zimmerman, B.J., & Martinez-Pons, M. (1992). Perceptions of efficacy and strategy use in the self-regulation of learning. In D. H. Schuck & J. L. Meece (Eds.), *Student perceptions in the classroom*. Hillsdale, NJ: Lawrence Erlbaum.

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 33-49

http://www.ijate.net

e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Ortaokul Matematik Öğretmenlerinin Dönütün Kullanımına İlişkin Görüşleri

Hacı Ömer BEYDOĞAN*

Ahi Evran Üniversitesi, Eğitim Fakültesi, Türkiye

Özet: Öğretim sürecinde dönüt, öğrencilerin akademik başarısını arttıran ve öğrencileri derse motive eden uyaranlardır. Dönütün öğretim sürecinde önemli bir yeri vardır. Dönüt öğrencilerin sadece öğrenme sürecine aktif katılımın sağlayan işlevsel bir uyaran değil; aynı zamanda akademik başarılarını da etkileyen bir araçtır. Bu araştırmada matematik öğretmenlerinin dönüt verme konusundaki görüşleri incelenmiştir. Çalışmada içerik analizi yöntemi kullanılmıştır. Çalışmanın verileri araştırmacılar tarafından oluşturulan yarı yapılandırılmış sorular ile yapılan görüşme tekniği ile elde edilmiştir. Görüşme sorularına öğretmenlerin verdikleri yanıtlar kaydedilip, içerik analizi yapılmıştır. Araştırmanın çalışma grubunu 2015–2016 öğretim yılında merkez ilçede 12 ayrı okulda görev yapan 12 matematik öğretmeni oluşturmuştur. Veriler matematik öğretmenleriyle yüz yüze görüşülerek elde edilmiştir. Tematik olarak incelenen öğretmen görüşleri, dönütte tarz, dönütte kapsam, dönütte ilke, dönütte yaşanan güçlükler, dönütte yetersizliğin kaynakları, dönütün faydaları, dönütün öğrenmedeki yeri temaları altında toplanmıştır. Sonuçlar öğretmenlerin, toplu öğretim sistemi içinde sunulan derslerde öğrencilerin matematik alanındaki boşluklarını doldurmanın mümkün olamayacağı şeklindedir. Elde edilen sonuçlar alan yazındaki verilerle ilişkilendirilerek tartışılmıştır. Elde edilen bulgular çerçevesinde bazı öneriler getirilmiştir.

MAKALE SÜRECİ

Gönderim: 21 Mart 2017
Düzeltme: 28 Temmuz 2017
Kabul: 14 Ağustos 2017

ANAHTAR KELIMELER

Dönüt, kapsamlı dönüt, Öğrenme süreci, Matematiksel kavramlaştırma, Matematiksel işlemler.

1. GİRİŞ

Dönüt, öğrencilerde öğrenmeyi artırmak amacıyla öğretmenlerin kullandığı önemli dışsal uyaranlardır. Öğretim sürecinde dışsal uyaranlara öğrencilerin yüklediği anlamlar sayesinde, öğretmen-öğrenci etkileşimi gerçekleşmektedir. Etkileşim sürecinde verilen dönütler, öğrencinin öğrenme gereksinimlerini karşıladığı; öğrenme çerçevesini uygun hale getirmek için kullandığı, öğrencilerin gelişim seviyelerine uygun olan sözel ve sözel olmayan uyaranlar haline getirildiğinde anlam kazanmaktadır (Looney, 2005). İşlevi gereği dönüt, öğrencinin, problem alanına odaklanması için etkileşim bilgisi sağlayan, bir sonraki işlem için harekete geçiren, anlayıp anlamadığını sorgulamasına fırsat veren, kendisini ve akranlarını değerlendirmesine hizmet eden uyaranlardır. Öğretmen dönüt vermeden önce öğrencinin öğrenmesi esnasında yaptığı yanlışları, özellikle kavram yanılgılarını belirlemesi ve bunları

*Corresponding Author E-mail: hobeydogan@gmail.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.339410**

belirlerken gözlem ve gözlem dışı teknikleri etkin kullanması gerekir. Bu durumun en etkin kullanıldığı değerlendirme türü, biçimlendirme yetiştirmeye dayalı değerlendirmedir (Kahl, 2005). Öğrenme ve öğretme sürecinde etkili olan dönütlerin öğretime pozitif katkıları olmakla birlikte dikkatsiz ve yanlış kullanıldığında da öğrenci öğrenmeleri üzerinde negatif etkiler oluşturmaktadır(Hattie &Timperley, 2007). Öğretim sürecinin başından sonuna kadar bütün süreçlerde yer alan dönütler, öğrencinin biçimlendirilip yetiştirilmesinde değerlendirmenin bel kemiğini oluşturmaktadır. Dönütler, öğretim sürecinde öğretmenler tarafından planlı bir şeklide kullanılabildiği gibi, zaman zaman kendiliğinden doğaçlama bir şekilde de gerçekleşebilir. Planlı olarak verilen dönütler, biçimlendirici değerlendirmenin en önemli öğesidir (Black & William, 1998; Hattie & Timperley, 2007). Dönütler öğrencinin bakış açısıyla incelendiğinde, öğrencinin kendi yaptıklarını doğruladığı, yanlışladığı veya önbilgisini değiştirmek amacıyla kullandığı, performansı ve anlayışı hakkında bilgi edindiği yardımcı uyaranlardır. Öğrenme sürecinde öğrencinin bu tür uyaranlara gereksinimi vardır (Mory, 2004).

Öğretmenler, öğretim sürecinde dönütleri, açık ve kapalı uçlu sorular şeklinde kullanabildiği gibi, öğrencinin yerine getirmesi gereken görev ve sorumlulukları yerine getirmek için anlık açıklamalar, öğrenmeyi kolaylaştırıcı eylem ve sözler şeklinde de gerçekleştirebilmektedirler. Dönütler farklı amaçlarla, öğrenmeyi etkilemek içinde kullanılmaktadır. Öğretmenler, öğrencilerde gerçekleştirilmesi öngörülen düzey ile öğrencinin gösterdiği performansı arasındaki boşluğu doldurmak veya gidermek; öğrencilerin öğrenme çabasını daha etkili hale getirmek; öğrenme-öğretme sürecinde öğrencilerin yanlış anlamalarını, işlem yanlışlarını ve kullandıkları strateji yanlışlarını düzeltmek için dönütü işe koşmaktadırlar.

Biçimlendirme yetiştirmeye yönelik değerlendirmede öğrencilerin başarısı hakkında alınan geri dönütler, öğretmen açısından öğrenme öğretme sürecini bütünleştirilmesi, öğrencilerin öğrendiklerini anlaması ve öğrencilerde öğrenmenin iyileştirilmesi noktasından önemli rol oynar. Değerlendirme sürecinde elde edilen veriler, öğrencilerdeki gelişimi ortaya koymakta, öğrenme ve öğretme sürecinde bir sonraki aşamada atılacak adımları, konu hakkında alınacak karaları oluşturmada açık ipuçları verir.

Dönütler, öğrencinin sadece akademik başarısını değil, aynı zamanda öğrencinin öğrenme sürecine etkin katılımını sağlamada (Brookhart, 2011) ve motivasyonunu sürdürmesinde etkin bir araçtır (Wigfield, Klauda & Cambria, 2008). Bu nedenle öğrencinin öğrenmeyi düzenleme yeterliğini de etkilemektedir. Klasik anlamda dönüt, öğrenciye dışarıdan sağlanan düzeltme bilgisi olarak tanımlansa da, çağdaş anlamda dönüt öğrencinin ne yapacağı ile ilgili bilgisini (Labuhn et al., 2010) çalışmasını adım adım iyileştirmesi için atacağı adımları ve çalışmalarının iyileşmesi için gereksinim duyduğu her şeyi içermektedir (Black & William, 1998; Hattie & Timperley, 2007; Sadler, 1998). Clark and Dwyer (1998), Foote (1999), Warden (2000), Zimmerman and Martinez-Pons (1992) dönütün öğrencilerin kavram yanılgılarını düzeltmeleri, bilgiyi zihinlerinde oluşturmaları, üst-bilişsel süreçlerini desteklemeleri, akademik performanslarını geliştirmeleri ve motivasyonlarını artırmaları noktasında yardım eden en önemli bilgi kaynağı olduğunu ifade etmektedir. Hattie ve Timberley (2007) göre dönütün esas amacı, öğrencinin hali hazırdaki anlayış ve performansı ile öğrenme hedefi arasındaki uyuşmazlığın vurgulanması; bu uyuşmazlığın azaltılması için öğrencileri cesaretlendirilmesı amacıyla atılan adımlardır (Rakoczy, Harks, Klieme, Blum & Hochweber, 2013). Bu bağlamda öğretmen ve öğrencinin akranlarından gelen dönütler, öğrencide performansını nasıl geliştirileceği (Butler & Winne,1995), sorumlu olduğu görevi nasıl yerine getireceği, öğrencinin gelişmelerini nasıl gözleyip değerlendireceği konusunda bilgi veren dışsal yönlendirmelerdir (Stone, 2000). Dönüt, öğrencinin hedeflenen duruma yönlendirilmesi için verilen açıklamalar veya uyaranlar olarak tanımlanabilir.

Öğretmenlerin sınıf içi dönüt düzeltme davranışlarıyla ilgili Türkiye'de yapılan araştırmalar incelendiğinde, yapılış yılları itibariyle araştırmalar şöyle sıralanabilir: Yunt, 1992; dönüt ve düzeltme etkenlerinin birlikte işe koşulmasının genel erişiyi önemli derecede yükselttiğini ortaya koymuştur. Saracaloğlu, Evin-Gencel ve Çengel (2011) tarafından yapılan araştırmada, öğretmen ve öğrenci görüşlerine göre lise öğretmenlerinin öğrenme ve öğretme süreci içindeki yeterlilikleri ele alınmış, öğretmenlerin sınıf içi davranışları konusunda kendilerini yeterli gördükleri halde öğrencilerin bunun tersi bir görüşte oldukları tespit edilmiştir. Şahin (2015), öğretmen adaylarının öğrenme ve öğretme sürecinde uygulanan dönüt etkinliği ile ilgili görüşlerini incelediği araştırmasında dönütü, açıklayıcı, eklemleyici, teşhis edici ve düzeltici dönütler olarak sınıflandırmıştır. Araştırmada öğretmenlerin teyit edici dönütü ara sıra, diğerlerini ise sıklıkla kullandıkları belirlenmiştir. Köğce ve Baki (2012) tarafından ilköğretim matematik öğretmenlerinin inanç ve davranışları üzerine yapılan bir araştırmada öğretmenlerin öğrenme-öğretme sürecinde dönüte genel olarak yer verdiği, ancak öğretmenlerin dönüt kavramına birbirinden farklı anlamlar yüklediği belirtilmiştir. Köğce ve Baki (2012), ilköğretim matematik öğretmenlerinin dönütü, öğrencilerin gösterdiği performanslarının yanında kişilik özelliklerine yönelikde bazı öğretmenlerin dönüt verirken olumsuz davranışlar sergiledikleri belirtilmiştir. Eraz ve Öksüz (2015) yaptıkları araştırmada, sınıf öğretmenlerinin ders dışı matematik etkinliklerinde verdikleri dönütlerin öğrencilerin başarı ve tutumlarına etkisini incelemiştir. Dönüt verilen gruplarda öğrenci başarısı ve olumlu tutum puanlarının diğer gruba göre anlamlı şekilde yükseldiğini ortaya koymuştur. Türkdoğan ve Baki (2012) de yaptıkları bir çalışmada, ilköğretim ikinci kademe matematik öğretmenlerinin yanlışlara yönelik verdikleri dönütleri gözlemleyerek, öğretmenlerin kullandıkları dönüt tekniklerini incelemişlerdir. Yukarıdaki çalışmalarda öğretmenlerin dönüt vermesinin öğrenci başarısına etkisi, kullandıkları dönüt türleri, dönüte yükledikleri anlam, öğrenci başarısı ve tutumuna etkisi üzerinde durulmuştur. Bu araştırmada, öğretmenlerin öğrencilere dönüt verirken, nasıl bir anlayış, ilke ve yaklaşımla dönüt verdiklerinin ortaya konması amaçlanmıştır.

1.1. Matematik Öğretiminde Dönüt

Öğrenciler tarafından soyut ve öğrenilmesi zor olarak algılanan matematik konularının kolaylaştırılması, öğrenciler tarafında anlaşılır hale getirilmesi için öğretme-öğrenme sürecinin işleyiş biçiminin dikkatlice gözlenmesi gerekir. Genel olarak öğrenilenlerin anlaşılır hale gelmesi için öğretmenlerin, öğretimin başında sonuna kadar öğrencileriyle çok yönlü etkileşime girmesi, uygun dönütler vermesi, matematiksel kavramların doğru anlaşılmasına ve sembolleştirilmesine yönelik atacakları adımlar, öğrenciler açısından son derece önemlidir. Matematik öğretiminde öğretmen dönütlerinin bu bağlamda izlenmesi ve eksiklerin belirlenmesi yoluyla gerçekçi çözümler üretilmesi bir gereksinim olarak ortaya çıkmaktadır. Matematik öğretiminde her öğrencinin, düşünme, muhakeme etme, düşündüklerini matematiksel sembollerle ifade edip dışarı vurması bir zorunluluk olarak ortaya çıkmaktadır. Bu nedenle matematik öğretiminde her öğrencinin başarısı, matematiksel sembolleri okuma, anlama ve uygulama yeterliği ile doğrusal ilişkilidir.

Uygulamada her öğretmenden, matematikle ilgili temel becerilerin kazandırılmasında duyarlılık göstermesi beklenir. Öğretmenler, öncelikli olarak öğrencileriyle bilişsel etkileşimi sağlamak, onları sorunun sebep ve sonucuna yönlendirmek için dönütü kullanmak zorundadırlar. Öğrencilerde matematik konularını öğrenirken attıkları her adımda ve yaptıkları her işlemde becerilerini iyileştirmek için dönüt alma gereksinimi duyarlar. Böyle bir süreci öğretmen, ancak öğrencileriyle etkileşim yoğunluğunu en üst düzeye çıkararak gerçekleştirebilir. Bunu yaparken, öğrencilerin öğrenme seviyelerini ve ön bilgilerini esas almak, uygun öğretim metotları geliştirip, farklı soru tipleri ile diyalogu artırmak, onların tahmin etme, analiz etme ve yorumlama yeterliklerini teşvik etme yollarını deneyebilirler

(Akyol, 2007). Bu özelliği ile dönüt, öğrencinin öğrenmesini iyileştirmesi, düşüncelerini ve davranışlarını değiştirmesi amacıyla sağlanan iletişim bilgisi olarak tanımlanabilir. Biçimlendirici dönüt öğrenciye sadece verilen bilgiyle değil, aynı zamanda öğrencinin öğrenmesini destekleyecek süreç ve aktiviteleri içermektedir.

Matematik öğretiminde, matematiksel kavramların kazandırılması sırasında oluşan kavram yanılgıları, işlem basamakları, işlemin tamamlanması ve işlemlerin yorumlanması sırasında yanlışların belirlenerek düzeltilmesine yönelik tepkilerin niteliği önem kazanmaktadır. Matematik öğretmenleri, öğretme-öğrenme sürecine yerleştirdikleri dönütlerle süreci eksiksiz tamamlamaya çalışırlar. Dönütün olmadığı bir süreçte öğrencinin matematik konularını öğrenmesi mümkün değildir. Çünkü dönütler, matematiksel düşünmenin ve kavramlaştırmanın, doğru işlem yapmanın omurgasını oluşturmaktadır.

Şantagata'ya (2002) göre matematik öğretim sürecinde öğretmenlerin dönütleri; Düzeltme, ipucu verme, soruyu tekrarlama, nedenini sorma, başka öğrenciye ipucu verme, soruyu dolaylı olarak tekrar sorma, doğru cevabı seçme, sınıftan doğru cevabı isteme, öğrencilerin girişimi ile doğru cevabı ortaya çıkarma şeklinde gerçekleşmektedir. Öğretim sürecinde genellikle öğretmenler, matematik öğretiminde yanlışları tespit ederek, yanlışlardan hareketle doğruyu bulma (sağlama yapma) yöntemi ile öğrencilere uygun beceriler kazandırabilir. Sınıf içinde öğrencilerin yanlışlarından hareketle anlamlı öğrenmeler ve beceriler kazandırabilirler (Nordstrom, Wendland and Williams, 1989). Çevikbaş ve Argün (2016) yaptıkları çalışmada, matematik öğretmenlerinin yanlış cevaplara verdikleri dönüt türlerinin öz saygı üzerinde hem olumlu hem de olumsuz etkilerinin bulunduğunu belirlemişlerdir.

Yurt dışında yapılan çalışmalarda ise Roschelle Rafanan, Bhanot, Estrella, Penuel, Nussbaum, Claro, (2010), yaptıkları araştırmada teknolojik destekli dönüt verilen grup çalışmalarının öğrencilerin matematik öğrenmeleri üzerine etkisini araştırmışlardır. Labuhn, Zimmerman ve Hasselhorn (2010) öğrencilerin özyeterlik algıları ve problem çözme performanslarının üzerine dönütün etkisi üzerinde dururken; Naroth, (2010) yapılandırılmış öğretmen dönütünün öğrencilerin matematik öğrenme performaslarının artırılması üzerine etkisini belirlemeye çalışmıştır. Carvalhoa, Santosa, Conboya and Martinsa (2014) 179 öğrenci üzerinde yaptıkları bir araştırmada öğrenci, algılarındaki farklılığın ortaya çıkarılmasında öğretmen dönütlerinin işlevi üzerinde durmuşlardır. Duhon, House, Hastings, Poncy and Solomon, (2015), zamanlaması ve açıklayıcılık özelliği açısından dönütün matematik öğremeye katkısını ele almışlardır.

Öğretme-öğrenme sürecinde dönütler, öğrencinin gereksinimiyle uyumlu verildiğinde, öğrencinin matematiği anlamasına, matematiksel sembolleri okumasına ve işlemleri doğru ilişkilendirmesine hizmet edebilir. Matematik öğretimine temel teşkil eden kavramlaştırma sürecinin etkin kullanımı öğrencinin öğrenmesine doğrudan etki eden bir durumdur. Öğretmenler bu tür dönütleri genellikle süreç içinde verirler. Matematiksel sembollerin doğru kullanılması ve ilişkilendirilmesi, aynı zamanda öğrencinin matematiksel içeriği anlaması ve bilişinde yapılandırmasında temel oluşturur.

Bu çalışmada, öğrencilerin matematik yeterliği edinmelerinde bu denli önemli işlevi olan dönütün öğretmenler tarafından nasıl algılandığını belirlemek, öğretmenlerin, dönüte yönelik düşünce ve eylemlerini tematik olarak incelenmesi ve alan yazında yer alan görüşler çerçevesinde değerlendirilmesi amaçlanmıştır.

Bu amaçla "Matematik öğretmenlerinin sınıf içinde verdikleri dönütlere ilişkin görüşleri nelerdir?" sorusuna yanıt aranmıştır. Bu problem cümlesi doğrultusunda aşağıdaki sorulara yanıt aranmıştır.

1- Size göre matematik öğretiminde dönütün işlevi" nedir

2- Öğretim sürecinde dönüt verirken nasıl bir yaklaşım izliyorsunuz?
3- Öğretmen olarak öğrenme öğretme sürecinde, öğrencilere genellikle hangi tür dönütleri hangi kapsamda veriyorsunuz.
4- Öğrencilere dönüt vermeniz gerektiğinde hangi temel ilkelerden hareket ediyorsunuz
5- Öğrencilere matematikle ilgili dönüt verirken zorlandığınız ve güçlük çektiğiniz noktalar nelerdir?
6- Matematik öğretmenleri için, öğretim sürecinde verilen dönütün faydaları nelerdir?

2. YÖNTEM

Bu çalışmada nitel araştırma yöntemi kullanılmıştır. Nitel araştırma, gözlem, görüşme ve doküman analizi gibi nitel veri toplama yöntemlerinin kullanıldığı, olguların ve olayların doğal ortamda gerçekçi ve bütüncül bir biçimde ortaya konmasına yönelik nitel bir sürecin izlendiği araştırma olarak tanımlanabilir (Yıldırım ve Şimşek, 2008). Çalışmanın verileri, ilköğretim ikinci kademede beş yıl ve daha fazla süre matematik öğretmenliği yapan 12 matematik öğretmeni ile gönüllülük esasına göre yüzyüze yapılan görüşmelerden elde edilmiştir.

Görüşmede, veri kayıplarını önlemek için kayıt cihazı kullanılmıştır. Katılımcılar kayıt cihazı kullanılacağı konusunda bilgilendirilmiştir. Yapılan görüşmelerin sonunda tutulan kayıtların katılımcılar tarafından dinlenebileceği, gerektiğinde kayıtlardaki görüşlerin isteğe bağlı olarak kısmen ya da tamamen çıkarılabileceği belirtilmiştir. Böylece kayıt cihazının katılımcılar üzerinde oluşturması muhtemel olumsuzluklar ortadan kadırılmıştır. Araştırma boyunca katılımcıların kendilerini rahat ve huzurlu hissetmeleri ve görüşlerini içtenlikle açıklamaları için bir görüşme ortamı sağlanmıştır. Görüşme sırasında, katılımcıların soruları cevaplarken araştırmacıdan etkilenmemesine çalışılmıştır. Araştırmada güvenirliği artırmak amacıyla öncelikle öğretmenlerin sınıf içinde kendi konumunu belirtmeleri istenmiştir. Veri kaynağı olan bireyler açıkça tanımlanmış, araştırma sürecinde oluşan sosyal ortamlar ve süreçler tanımlanmıştır.

Görüşmede katılımcılara araştırmacı tarafından geliştirilen yarı yapılandırılmış 6 soru yöneltilmiş, katılımcıların verdikleri yanıtlar kaydedilmiştir. Katılımcılarla yüzyüze yapılan görüşmelerde katılımcılara bir önceki görüşmede sorulan soruların bir kısmı tekrar sorularak benzer yanıtlar verip vermedikleri kontrol edilmiş, toplanan bilgilerde tutarlılık sağlanmaya çalışılmıştır. Tutarlık gösteren ifadeler analiz sürecine dâhil edilmiştir. Katılımcılarla yapılan görüşmeler 4 haftalık bir süreyi kapsamıştır. Görüşmede kaydedilen sözel açıklamalar, deşifre edilerek çözümlenip yazıya geçirilmiştir. Görüşmede elde edilen içerik tematik olarak analiz edilmiştir.

Çalışma grubunu, 2015–2016 öğretim yılında merkez ilçede yer alan beş yıldan fazla öğretim deneyimi olan 12 ayrı okulda görev yapan 12 matematik öğretmeni oluşturmuş ve görüşmelere gönüllü olarak katılmışlardır. Görüşmelerde elde edilen veriler, kaydedilip, daha sonra çözümlenerek yazılı hale getirilmiştir. Elde edilen sonuçlar alanyazındaki verilerle ilişkilendirilip, tartışılarak sunulmuştur.

2.1. Analiz süreci

Matematik öğtmenlerinin görüşlerinin analizinde, ifadelerin benzerliğine göre gruplamalar yapılmış; çözümlemede görüşüne başvurulan öğretmenlere birer kod numarası verilerek (K1, K2…) açıklamalar yapılmıştır. İfadelerdeki benzer öğeler gruplandırılmış ve gruba uygun olarak temalar adlandırılmıştır. Temalardaki tutarlılığı sağlamak için temaları oluşturan kavramlar kendi aralarında gruplandırılmış, temaların diğer temalarla tutarlılığı değerlendirilmiş ve anlamlı bir bütün oluşturup oluşturmadığı test edilmiştir. Bulguların daha önce yapılan araştırmalarla uygunluğu karşılaştırılmıştır. Temalar, duruma göre tümdengelim

ya da tümevarım yöntemi ile açıklanmış ve yorumlanmıştır. Bulgular, katılımcılar tarafından gözden geçirilmiş ve gerçekçi bulunmuştur.

Araştırma bulgularının öngörülerle tutarlığı esas alınmıştır. Elde edilen verilerde dış geçerliği sağlamak için; Veri toplama aracının hazırlanmasından, uygulama ve analiz aşamasına kadar geçen araştırma sürecinin detayları açıklanmıştır. Bulgular, alan yazınla karşılaştırılarak, bulguların uygulamadaki gerçekliklere uygunluğu belirlenmeye çalışılmıştır. Araştırmanın başka araştırmalarla test edilebilmesi için gerekli açıklamalar ayrıntılı olarak betimlenmeye çalışılmıştır. Görüşleri alınan katılımcılarla (matematik öğretmenleryile) tekrar görüşülmüş ve bulgular paylaşılarak doğrulanmıştır.

3. BULGULAR

Araştırmada elde edilen bulgular çerçevesinde temalar ortaya çıkarılmıştır. Temalar, dönütün öğretimdeki yeri, dönütün veriliş tarzı, dönütün kapsamı, dönütte ilkeli davranma, dönütte yaşanan güçlükler ve dönütün faydaları gibi başlıklar altında toplanmıştır. *Dönütün öğretimdeki yeri teması* altında aşağıdaki noktalar ön plana çıkmıştır:

Belirsizliği kaldırma	K12
Eski bilgi ile yeniyi harmanlama	K11
Kalıcılığı sağlama	K1, K11
Hatayı azaltma	K4
Anlayışı kolaylaştırma	K7, K12

Öğretim sürecinde dönüt, öğretmenlerin öğrencilerde öğrenmeyi artırmak amacıyla kullandığı dışsal uyaranlardır. Öğrenciler dışsal uyaranlara yükledikleri anlama uygun etkileşimlerini sürdürürler. Etkileşim sürecinde verilen dönütler, öğrencinin öğrenme gereksinimlerini karşıladığı; öğrenme çerçevesini uygun hale getirmek için kullanıldığı, öğrencilerin gelişim seviyelerine uygun sözel ve sözel olmayan uyaranlar haline getirildiğinde anlam kazanmaktadır (Looney, 2005). İşlevi gereği dönüt, öğrenciye gerekli bilgiyi sağlayan, bir sonraki işlem için harekete geçiren, anlayıp anlamadığını sorgulamasına fırsat veren, kendisini ve akranlarını değerlendirmesine hizmet eden uyaranlardır.

Öğretmenler, öğrenme sürecini öğrencileriyle birlikte planlar ve öğretim hedeflerini oluşturur. Öğrencilerin öğrenmeleri sırasında ortaya çıkan eksikleri ve boşlukları zamanında fark eder; boşlukları ortadan kaldıracak şekilde dönütler verir. Öğretmenler, öğrencilerin öğrenmelerinde ortaya çıkan boşlukları doldurmak için farklı dönüt verme stratejilerini kullanırlar. Matematik öğretmenlerine, "Size göre matematik öğretiminde dönütün işlevi" nedir? Şeklindeki açık uçlu soru yöneltilmiş, öğretmenlerin verdikleri yanıtlar, aşağıda verilmiştir:

K12: Dönüt, öğrencilerin matematik konularını daha iyi anlaması için verilir. Öğrencinin kafasını karıştıran herhangi bir durum olduğunda öğrenci öğretmene sorması ve yanıtını almasıdır. Dönüt, öğrencinin kafasındaki soru işaretlerini kaldırmaya yarar. Verilmediği zaman öğrenci konuyu tam olarak anlayamaz, bir sonraki hafta konu anlatıldığında, öğrencinin aklında hiç bir şey kalmaz. Zaman geçtikçe öğretmenin o öğrenciyi toparlaması gittikçe zorlaşır.

K11: Dönütü özellikle sınavlardan sonra veriyoruz, çocuk eski bilgisi ile yeni bilgisini harmanlayabilsin diye. O yüzden dönüt vermek gerçekten eğitimin olmazsa olmazıdır. Öğrendikleri davranışların kalıcı olup olmadığını yeni davranışları edinip edinmediğini görmemiz açısından önemlidir.

K1 : Öğrenci yanlışta yapsa, doğru da yapsa ancak uğraştığında bir şeyler kalır, kendisinde. Birşeyler kaldığında, kalan şeylerin daha değerli olduğunu hisseder.

K4 : Öğrencinin anlamadığı noktaları aydınlatacak dönütler verdiğimizde öğrenci çok faydasını görüyor. Çocuk aynı hataya bir daha düşmüyor, aynı hatayı tekrar yapmıyor.

K7 : Öğrencilerin, matematik konularını yanlış anlamasını, yanlış çıkarımda bulunmasını ve yanlış sonuca gitmesini engellemek için derslerde gerektiği kadar açıklamalar yapıyoruz. Buna ragmen öğrencide sayısal konulara yatkınlık yok ise istediğimiz sonucu alamıyoruz.

Öğretmenler, öğretme-öğrenme sürecinde verilen dönütlerin, matematik içeriğinin iyi anlaşılması, öğrencinin önbilgileri ile yeni bilgilerini bütünleştirmesi, öğrencilerin zihinindeki çelişkinin kaldırılması, öğrenmelerde kalıcılığın artırılması, öğrencinin gereksinim duyduğu bilgi ve becerilerin kazandırılmasına katkı getirdiğini dile getirmektedirler. *Dönütte tarz teması* altında aşağıdaki noktalar ön plana çıkmıştır:

Sonuca odaklama	K5
Yanlışı gösterme	K5, K8
Analojiden faydalanma	K8
Ben dili kullanma	K8, K2
Birlikte sorgulama	K11, K5, K7
Birebir ilgilenme	K12, K1
Süreci ve ve sonucu kontrol	K12, K2, K5, K1

Etkili bir iletişimde paylaşılan mesajların, hem öğrencinin hem de öğretmenin mesajın içeriğini doğru anlamaları açısından önemlidir. Mesajın açıklık kazanması ise dönütün veriliş tarzıyla yakınen ilişkilidir. Öğretmenlere, öğretim sürecinde dönüt verirken nasıl bir yaklaşım izliyorsunuz? Sorusuna verdikleri yanıtlar şu şekildedir:

K5 : Öğrenciye cevap verirken yanlış yapmışsın şeklinde değil de, burasını böyle çözdüğünde yanlış sonuca ulaşıyorsun, bunu o şekilde değil de, şu şekilde çözersen istenilen sonuca ulaşabilirsin şeklinde sonuç odaklı dönüt veriyorum.

K8 : Burası olmamış nasıl yaptın böyle demek yerine, şurası doğru ama, şuradan sonrası yanlış diyorum. Yanlıştan hareketle doğruyu bulması için öğrencileri tartıştırıyorum. Sınıfta benzer yanlışları yapan öğrenciler içinde iyi bir uygulama oluyor. Öğrenciler yanlış yapmaktan değil, yanlışı düzeltmekten ve sınıf içinde tartışmaktan keyif alıyor.

K11:Matematikte dönütü çocuklara bireysel olarak vermeyi yeğliyorum. Çünkü her öğrencinin takıldığı nokta, birbirinden farklı oluyor. Her öğrenci çözümünü bana tek tek getiriyor ve o sırada takıldığı noktayı gördüğümde ha bak, çözümün şu noktasına tekrar bakman gerekir dediğim zaman hani direkt onun hatasına yönelik açıklama yaptığım için o, çocukta daha çok yararlı oluyor.

K12:*Öğrenciyi kırmamak gerekir. Mesela bazı öğrenciye üç sefer anlatıyorum, öğrenci artık ben bu işi yapamam, diyemiyor, ha gayret evladım, aferin güzel olmuş diyorum ondan sonra bu çocuğun üzerindeki yapamam baskısı da azalıyor. Başka bir seferinde diyorsun evladım bak bunu gerçekten böyle yapman gerektiğini görüyorsun değil mi diyorum. Ondan sonra derse biraz daha ilgisi artıyor. Kısacası dönütün verildiği ortam ve dönütün veriliş biçimi çok önemli.*

K5: Dönüt verirken, yani çocuğu aşağılayarak, bağırarak uyarmak var, bir de tatlı dille uyarmak var. Ne kadar tatlı dille uyarırsam o kadar iyi sonuç aldığımı düşünüyorum. Ne demişler tatlı dil, yılanı deliğinden çıkarır. Basit sorular sorarak çocuğu sorunun çözümüne odaklarım. Gerektiğinde kuralı hatırlatıcı ipuçları veririm, öğrencilerin öğrenme gereksinimlerini gidermeye çalışırım.

K7: Çocuğun bulduğu sonuçlar çözümle örtüşmediğinde önce durum tespiti yaparım. Bu sonucu nasıl buldun diye sorarım ve öğrenciden durumu açıklamasını isterim. Eğer öğrenci gerçekten doğru işlem yapmışsa sınıf huzurunda teşekkür ederim. Eğer yanlış çıkarım, yanlış işlem yapmışsa, sorunu öğrenciyle birlikte sorgulayarak problemi çözdürürüm.

K2: Çocuklar problemleri farklı yollardan da çözüyorlar. Mesela, oran, orantı sorusunda bügün yaşadım, çocuk farklı yerlere yazmış ama sonuç aynı çıkıyor değişen bir şey yok yani. Öğretmenim ben, farklı yoldan buldum bakabilir misiniz? Diyor. Hemen gidip, kontrol

ediyorum, anında ona cevap veriyorum, hoşuna gidiyor, ilgisi artıyor, şevki artıyor. Kısacası çocuklar özel ilgi istiyor.

K1: Etkinlik yaparken sürekli öğrencilerin yanına giderek dönüt veririm, çünkü ben onların yanına gittiğimde ilgilendiğimi iliklerine kadar hissederler.

K2: Vücut dilimi kullanırım. Ben çok bağırmam, mesela şu el hareketini (ellerimi göğsümün üstünde üst üste bağladığımda) yaptığım zaman hepsi yaslanır eller arkada. Kızmam da, bakın der, açıklamamı yaparım.

K5: Sınıflarda U masası düzeni çok önemli. Yani çocukların U şeklinde oturmaları, birbirlerini görmeleri gerekir, bütün sınıflarda öğrenciler bu şekilde oturmuş olsa, herşeyden once çocuklar birbirini kontrol eder, etkileşime girer ve çocuklar birbirinden öğrenir.

K5: Sınıf içinde öğrencilerin problemin çözümüyle ilgili pek soruları olmuyor. Göz ucuyla alıştırmalarını kontrol ediyorum ancak öğrenciyi sınıfta tahtaya kaldırarak soruları çözdürdüğümde öğrenciden geri dönüt alabiliyorum.

K1: Sınav sisteminin mutlaka değişmesi gerekir, öğrenciler sınava yönelik hazırlanmamalı, öğrenciler temel bilgi ve becerilere sahip olacak şekilde öğrenmeli, onun içinde etkinliklerin arttırılması lazım. Sınav kaygısı ve baskısı matematik öğretmenini derslerde sadece soru çözen bir insan haline getiriyor, Maalesef !..

Öğretmenler, öğrencilere verilen dönütlerde öğrencilerin öğrenme gereksinimine odaklanılması gerektiğinde hemfikirler. Dönütün öğrencinin doğrudan eksiğine, yanlışına, anlayışına ve yetersizliğine yöneltilmesi halinde öğrencinin bu etkileşimden hoşlandığının ve daha fazla faydalandığının bilincindeler. Matematikte dönütlerin, öğrenciye açık seçik iletilmesi, çoklu etkileşimin sağlanması için öğrencilerin oturma şeklinin önemine vurgu yapmaktadırlar. Matematik derslerinde sadece öğretmenin aktif ve problem çözen kişi olması yerine, öğrenci merkezli etkinliklere dayalı bir matematik öğretimiyle öğretiminde başarılı olabileceği görüşündeler. *Dönütte kapsam teması altında aşağıdaki noktalar ön plana çıkmıştır:*

Açıklama, gösterme ve çözdürme	K7, K5
Düşündürme	K2
Benzerliklerden ve zıtlıktan faydalanma	K5
Ayrıntıları vurgulama	K11
Zaman ayırma	K3

Öğretmen olarak öğrenme öğretme sürecinde, öğrencilere genellikle hangi tür dönütleri ne düzeyde veriyorsunuz? sorusuna öğretmenlerin verdikleri yanıtlar şu şekildedir:

K7: Öğrencinin kafasında hiç bir soru işareti kalmaması için öğreteceğim konuyu açıklarım, gösteririm, çözerim ve yeteri kadar örnek veririm. Gerekirse ekstra fazladan örnekler çözerim ya da öğrencinin kendisine çözdürürüm. Bunları yapmassak çocukların matematiği öğrenmeleri çok zor.

K2: Toplamayı, çıkarmayı ve matematiği ezberletirseniz çocuk biter. Matematik ezberletilmez, çocuğun biraz soyut düşünmesi lazım. Matematik kavramlarını doğru anlaması, uygun sembolleri ilişkilendirmesi ve içselleştirmesi çok önemlidir.

K5: Öğrencinin doğru çıkarım ve doğru çözüm yapması için zihninde bulanıklığın olmaması gerekir diye düşünürüm. Yanlış işlem ve yanlış çıkarım, o çocuğun zihnindeki eksik bilgi ya da yanlış algı sonucu oluşmaktadır. Bir sorunun mümkünse farklı yollarla çözmünü gösterir, gerekirse sağlamasını yaparım. Öğrencilerden de benzer yolları denemelerini isterim. Zihninde bulanıklık olmayan öğrenci yanlış yapmaz, çıkarımda da bulunmaz.

K3: Öğrencinin seviyesi çok düşük olduğunda, yani temel becerilerde eksiklikleri olduğu zaman onun üzerine ne kadar ne koyabilirsiniz? Konuyu ne kadar açıklarsanız, açıklayın, çocuk

anlayacak seviyede olmadığı için anlamıyor. Aynı zamanda doğrusu bizde, öyle pek üzerinde duramıyoruz. Çünkü üzerinde durduğumuz zaman dersin belki yarım saati, bir saati gidiyor, o çocuğa zaman ayırdığımızda, öbür çocukları ihmal ediyoruz.

K11: Bu sene, ben 5. Sınıf matematik derslerine giriyorum. Derslerde, bütün etkinliklerin çözümlerini tek tek irdeleyerek yaptırıyorum, çocukların sorularına da cevap veriyorum. Zaman problemim yok, çünkü 5 saat dersim var 2 saat seçmelim var ve 4 saatte kursum var bu kadar sürede her öğrenciye ayrıntılı tek tek cevap verme fırsatım oluyor.

Matematik konuları içeriği itibariyle soyut, muhakeme ve çıkarım gerektirmektedir. Bu dersin basmakalıp olarak ezbere öğretilemeyeceği, öğrencinin etkinliklerden hareketle eksiklerini giderici dönütlere gereksinim duyulduğu açıkça ortadadır. Matematik öğretiminin kavramlardan, işlemlere doğru ilerleyen aşamalı bir süreci içerdiği dile getirilmektedir. Öğrencinin zihninde hem kavramsal açıdan, hem de işlemsel açıdan eksiklerinin giderilmesi için kapsamlı dönüt verilmesi gerektiği öğretmenler tarafından önemsenmektedir. *Dönütte ilkeli hareket etme teması altında aşağıdaki noktalar ön plana çıkmıştır:*

Var olan yeterliği ön plana çıkarma	K8, K3
Basit örneklerden hareket	K4
Ben dili kullanma	K1, K11
Yanlışı gösterme	K1, K7, K11
Motive etmede mevcut doğruları kullanma	K3
Bireysel iletişim kurma	K12
Zamanlama	K5

Matematik öğretirken öğrencilere dönüt vermeniz gerektiğinde hangi temel ilkelerden hareket edersiniz? Sorusuna öğretmenlerin verdikleri yanıtlar şu noktalarda toplanmıştır:

K8: Öğrencinin küçük bir doğrusu varsa bile, onu ön plana çıkarıp aferin, bak burayı doğru yapmışsın, ama devamında buralar böyle olmayacak, şöyle olacak şeklinde söylemek, çocuğun özgüvenini arttırıyor.

K11: Yaptığım açıklamaların açık anlaşılır olmasına ve sakin sakin, kızmadan yavaş yavaş çocuğun anlayacağı şekilde sunmaya dikkat ederim. Seviyesi düşük olan öğrencilere daha basit örnekler veririm, verdiğim örneğide biraz daha yavaş sunmaya gayret ederim.

K1: Sert bir şekilde, sen niçin yanlış yaptın, öyle değil, bu şekilde, şu şekilde… işte o cümleleri kurduğumuz zaman zaten öğrenci üzerindeki başarı giderek düşüyor yani.

K7: Öğrenci yanlış yaptığında, anlamadığı yerleri farketmesi için onu tahtaya kaldırmak, yanlışlarını göstermek, bak bunu böyle yaptığında yanlış, ama şöyle yaptığında doğru sonuca ulaşıyorsun diyorum. Öğrenci kontrolümde doğru çözüm yaptıktan sonra bir örnekte kendin çöz diyorum.

K3: Çocuk işlem yaparken doğruları varsa şuralar doğru dediğiniz zaman çocuğun kendine güveni geliyor ve faydalı oluyor. Onun için çocuğun doğrularını söylüyorum.

K12: Yanlış davranışların düzeltilmesinde veyahut da doğru davranışların pekiştirilmesinde öğretmenin ilk yapması gereken şey öğrencisinin adıyla, soyadıyla, kişiliğiyle, yapabilecekleriyle ve yapamayacaklarıyla tanımalı. Eğer, öğretmen gerçekten öğrencisini tanıyorsa, öğrencisini her zaman değerli bir insan olduğunu hissettirmesi gerekir. Bu durum başarıyı artırır; yoksa dönütleriniz, boş konuşulan, boşa çekilen kürek misali, anlamsızdır. Bildiklerinizi anlatırsınız o yetmez. Siz o öğrenciyi kazanmalısınız. Öğrenciyi tanımadığınızda konuyu açıklarken şişt... oğlum kızım diyerek, maalesef yardımcı olamazsınız. Ama ailesini tanıyorsanız, ailesinin telefonu sizde kayıtlı ise, böyle bir durumda hemen aileyi işin içine katabilirsiniz aileyle diyalog halinde olursunuz ki o zaman öğrenci benim kaçacak yerim yok der.

K5: Öğrenciye dönüt verirken zamanlamanın çok uygun olması gerekir.

K4: Önce öğrencinin konu hakkında bir önbilgisini ölçmek lazım, aslında. Öğrencide o ön bilgi yoksa öğretmen olarak senin o ön bilgiyi çocuğa vermen lazım.

Yapılan araştırmalardan, elde edinilen bilgi, deneyim ve gözlemlerden hareketle bir takım genel doğrulara ve çıkarımlara ulaşılması gerekir. Öğretmenlerin dönüt verirken ilkelere ne kadar uyduklarını belirlemek için görüşleri incelendiğinde, ortaya çıkan görüşler şu şekilde betimlenmiştir: Öğretmenler öğrencilerini farklı tekniklerle tanımaları gerektiğini önemsiyor. Öğrenciyi yüreklendirmek için önce yapabildiklerinden ve yeterliklerinden bahsedip, sonra eksiğine yer vermenin duygusal olarak öğrencinin dengelenmesi ve özgüven oluşturmasına katkı getireceği kanatindeler. Bunu gerçekleştirirken de sert, suçlayıcı dil kullanmak yerine, "ben dili "içeren yumuşak bir üslubun benimsenmesi gerektiği noktasında ortak görüşe sahipler. Öğrencilerin önce verilenleri almaya hazır hale getirilmesi, verilen dönütlerin düzeyine uygun seçilmesi, örneklerle somutlaştırılarak sunulması, genel öğrenme ilkelerine bağlı kalınması gerektiğini vurgulamaktadırlar. *Dönüt vermede yaşanan güçlükler* teması altında aşağıdaki noktalar ön plana çıkmıştır.

Önbilgi eksiği	*K11, K1*
Öğrenci seviyesine inememe	*K11, K1, K8*
Öğrenciye yeteri kadar zaman ayıramama	*K11, K10*
Algıda, anlamada ve çözümde zorlanma	*K9, K1*
Aşırı hereketlilik ve öğrenmede direnç	*K1, K8*
İçerik yoğunluğu	*K12*

Öğretme-öğrenme sürecinde öğretmenlerin, "öğrenciye matematikle ilgili dönüt verirken zorlandıkları ve güçlük çektikleri noktalar" a ilişkin görüşleri şöyledir:

K5: Öğrencilerin konuya ilişkin ön bilgilerinde eksiklik olduğu zaman onlara cevap vermekte güçlük çekiyorum.

K9: Dönüt öğrenci seviyesine bağlı bir şey, verilen dönütü seviyesi iyi olan öğrenciler alıyor, özellikle orta seviyedeki öğrencilerde dönüt verme ve alma karşılık oluyor. Seviyesi düşük öğrencilerde o bağlantıyı kurmak çok zor oluyor.

K10:Sınıfta her öğrenciye dönüt vermek çok zor çünkü sınıflarımızda öğrenci sayısı çok fazla, bazı öğrencilerimizde ilgi çok düşük. Her çocuğun önce soruyu algılayıp sonra çözüme geçmesi gerekir. Biz çözümleyemediği noktada ona yardımcı olmaktayız. Ama çocuk zaten soruyu çözmek istemiyor. Anlamakta da direnç gösteriyor. Bu noktada dönüt vermek mümkün değil. Dönüt nedir bir şeyi atarsın o da atılanı geri karşılıklı olarak sana atar. Yani karşılıklı etkileşime girersin. Tek yönlü etkileşimden sonuç alınmıyor.

K1: Öğrenciler hareketli, içeride dışarıda sürekli birbirleri ile itişip, kakışyorlar ve didişme halindeler. Yani bu bir konuşma olsun, bu bir soruna ilişkin açıklama olsun, bu bir arkadaşlık olsun farketmiyor onlar için. Grup içinde verilen bir dönütü öğrenmede direnç davranışına çeviriyorlar. Böyle bir fırsat ellerine geçtiği anda onu hemen değerlendiriyorlar "baksana yapamamış", onu da çözememiş gibi falan diye.

K11:Öğrencilerin temeli zayıfsa ben ne anlatırsam anlatayım ne kadar basitte indirgersem indirgeyeyim yine de anlamadığı oluyor. Bu öğrenciler için belki 3 sene 4 sene geriye gitmek gerekiyor. Ama bunu yapmak çokta kolay değil, işte bu çok zor maalesef. Mesela 5. sınıftan birkaç çocuk daha 1. 2. sınıf seviyesinde. Farkındayım onlar da çok zorluk çekiyor… Onlarla problemi ancak şöyle aşmaya çalışıyorum. Sınıfta çalışkan bir arkadaşına bu konuyu anlatabilir misin diyorum? Verilen görev, çalışkan öğrenci içinde yararlı oluyor anlatınca kendiside öğrendiklerini iyice pekiştiriyor. Diğeri içinde faydalı oluyor, çünkü ben he öğrenciye yeteri kadar zaman ayıramıyorum. Eğer ailede abi varsa abla varsa o şekilde görevlendirerek ya da veliyle konuşarak durumu idare ediyorum.

K12:Öğrencilere yeteri kadar dönüt vermede zorlanıyorum. Çünkü işlememiz gereken o kadar çok konu var ki, müfredatı yetiştirmek durumundayım. Deneme sınavları geldiği zaman konuları tekrar etmek durumundayım. Büyük bir yarış ortamında çalışıyoruz. Okul, şehrin

derece yapan okulu. Müdür Bey, benim sınıfımın başarı ortalamasının % 94 olduğunu biliyor. Sınıfımın başarı ortalaması % 94 ten % 92 ye düşse, sınıfınızın başarısı düşmüş hocam diyor, sınıfın sorumlusu sizsiniz diyor. Her öğrenciye ayrıntılı dönüt vermek yerine toplu bir şekilde dönüt vererek hem zamandan kazanmaya hem de başarımı sürdürmeye çalışıyorum. Aynı zamanda ğrencilere bazı yanlış davranışlarını gösteriyorum. Çocuklar yanlışlarını fark ederek, doğrusunu sizden görerek ancak sonuca ulaşabiliyor.

K8: Verici ve alıcı arasındaki ilişkinin sağlıklı olabilmesi için, alıcının nerde ne kadar mesaj alabileceğini iyi bilmemize bağlı. Bir çocuğun altsınıflarda öğrenmesi gereken temel bilgileri öğrenmeden yeni bilgileri kavraması mümkün değil. Bundan dolayı dönüt verirken bazen düzeyi aşağılara kadar çektiğim oluyor. Bu seferde sınıftaki diğer öğrenciler, o öğrenciyle dalga geçiyor veyahut da o sınıfta iletişim o noktada kopuyor. Mesela sınıfın bir tarafına basit, temel bir bilgiyi sunuyorsunuz ama diğer taraftan temel bilgiyi sahip olan çocuklar, sahip olmayan çocuklara karşı tepki veriyorlar. Ya bunu da mı bilmiyorsun diyor, bu sefer anlamayan öğrenci daha çok içine kapanıyor. Bu sefer o öğrenciden geriye dönüt almakta iyice zorlaşıyor.

K1: Seviyesi düşük öğrencinin üzerinde çok durmak ders ortamında pek mümkün olmuyor. Dışarıda teneffüste veya boş zamanlarımızda onun üzerinde durulabililiyoruz.

Matematik öğretmenlerinin, dönüt verirken öğrencilerin matematikle ilgili hazır bulunuşluk düzeylerinin zayıf olması, konuya ilişkin ön-yaşantılarını eksikliği, öğrenme seviyelerinin düşük olması, matematik derslerindeki içerik yoğunluğu ve matematik konularının soyutluğuna bağlı olarak öğrencilerde başaramama korkusu ve özgüven eksikliği, sınıfların kalabalık olması, her öğrenciye yeteri kadar zaman ayıramama noktasında zorlandıkları anlaşılmaktadır: *Dönütün Faydaları* teması altında aşağıdaki noktalar ön plana çıkmıştır:

Farkına varma	K9, K2, K7, K5
Ilgide artış	K9
Olumlu tutum	K8, K7
Özgüven	K9
Kendini kontrol	K5

Matematik öğretmenlerine, öğretim sürecinde dönütün faydaları sizce nelerdir? şeklinde yöneltilen soruya öğretmnelerin verdikleri cevaplar şöyledir:

K9: Öğrencilere dönüt vermek önemli, çünkü öğrenci yaptığı işlemin sonucunun doğruluğunu veya yanlışlığını bireysel olarak anladığında öğrenci şunu fark ediyor, öğretmen benimle ilgileniyor. Öğrencinin derse ilgisi artıyor.

K2: Öğrencilere bireysel dönüt vermek önemlidir. Çocuğun sorduğu veya takıldığı bir şeyi gördüğümde ha bak şuraya dikkatli bak dediğim zaman, direk onun hatasına yönelik konuştuğum için o çocuk için bu yanıt, daha yararlı oluyor.

K8: Dönüt, öğrencinin diğer derslere katılması, kendine özgüveni, derse ve öğretmene yaklaşımı açılarından bence çok faydalıdır. Fırsat bulduğumuz ölçüde öğrenciye yanlışlarını ve doğrularını görecek şekilde dönüt vermemiz gerekir.

K5: Öğrencilere dönüt verirken her şeyden once biz kendimizi kontrol ediyoruz. Yani derslerde neyi ne kadar verdik, bu öğrenci ne kadar anladı, anlaşılmayan yerleri görme ve tekrar üzerinde durma şansı elde ediyoruz.

K7: Dönüt, öğrencinin tam olarak konuyu anlamasına yardımcı oluyor. Kafasında kalan soru işaretlerini ortadan kaldırıyor.

K2: Nerede hata yapmış? Acaba bu hatanın kaynağı ne? Acaba matematikte toplama çıkarma çarpma bölme işlemini bilmiyor mu? Formülü mü bilmiyor? Konuyu mu hiç anlamamış? Bunlar bizim için çok önemli. Çünkü dönüt bizim için bir anahtardır.

Öğretmenler, öğrencinin anlatılan konuyu ne kadar anlayıp veya anlamadığını görmek, anlamadığı konuları tekrar etmek, öğrencilerin eksiklerini doğrudan öğrencinin yüzüne söylemenin daha doğru olduğunu düşünmektedir. Öğretmenler, dönütün, öğretmen olarak kendilerini değerlendirmelerinde, öğrencilerinde özgüven oluşturmalarında, öğrencilerinin öğretime katılımlarını sağlamada, öğrencilerin öğrenmeye odaklamasını sağlamada faydalı olduğu nokasında hem fikirler.

4. TARTIŞMA VE SONUÇ

Matematik öğretmenleriyle yapılan görüşmelerde, öğretmenlerin sınıf içinde verdikleri dönütün yapısı tematik olarak ele alınıp incelenmiştir. Temalar, dönütün öğretimdeki yeri, dönütün veriliş tarzı, dönütün kapsamı, dönütte ilkeli davranma, dönüt vermede yaşanan güçlükler ve dönütün faydaları başlıkları altında toplanmıştır.

Araştırmaya katılan öğretmenlerin görüşleri doğrultusunda dönütün öğretimdeki yeri, öğrenmede belirsizliği ortadan kaldırma, eski bilgi ile yeni bilgiyi harmanlama, kalıcılığı sağlama, hatayı azaltma ve anlayışı kolaylaştırma noktalarında toplanmaktadır. Dönütün öğretimdeki yeri konusunda öğretmenlerin büyük bir kısmı bilinçli, öğrencilerine yeteri kadar kapsamlı ve uygun dönüt verememenin endişesi içindedirler. Bu endişelerini sınıfların kalabalık olması, öğrencilerin hazırbulunuşluk düzeylerinin düşük olması gibi gerekçelerle açıklamaya çalışmaktadırlar. Matematik öğretiminde öğretmen, dönütleri, yanlışı düzeltme, ipucu verme, soruyu tekrarlama, nedenini sorma, başka öğrenciye ipucu verme, soruyu dolaylı olarak tekrar sorma, doğru cevabı seçme, sınıftan doğru cevabı isteme, öğrencilerin girişimi ile doğru cevabı ortaya çıkarma şeklinde gerçekleştirmektedir (Şantagata, 2002). Bazı durumlarda da öğretmenler, sınıf içinde öğrencilerin yanlışlarından hareketle anlamlı öğrenmeler ve beceriler kazandırmaktadırlar (Nordstrom, Wendland & Williams, 1989). Köğce ve Baki (2012) ilköğretim matematik öğretmenlerinin dönütü, öğrencilerin gösterdikleri performansın yanında kişilik özelliklerine yönelik kullandıklarını; Eraz ve Öksüz (2015) yaptıkları araştırmada dönüt verilen gruplarda öğrenci başarısı ve olumlu tutum puanlarının diğer gruba göre anlamlı şekilde yükseldiğini ortaya koymuşlardır.

Matematik öğretmenleri dönütü öğrencinin gereksinim duyduğu durumda gerektiği kadar ve kapsamlıca verdiklerini belirtmektedir. Şahin (2015) öğretmenlerin dönütü, öğrencinin dikkatini çekme, güdüleme, hedeften haberdar etme, işaret ve ipuçlarını sunma ve önkoşul davranışları kazandırma etkinliklerinde sıkça kullandıklarını ortaya koymaktadır.

Dönütün veriliş tarzına ilişkin temada, matematik öğretmenlerinin görüşleri, öğrenciyi sonuca odaklama, yanlışı gösterme, analojiden faydalanma, ben dili kullanma, birlikte sorgulama, birebir ilgilenme, süreci ve ve sonucu kontrol etme noktalarında toplanmaktadır. Öğretmenlerin neden dönüt verdiklerini ve sorunun ne olduğunu belirlemek için bir çaba içine girdikleri görülmektedir. Öğretmenlerin, öğrencilerin konuyu bilip bilmediğini anlamak için basit sorular sorması, sorunun çözümü ile ilgili kuralları hatırlatmaları, çözüm yoluna dönük ipuçları vermeleri, öğrencilerin öğrenme gereksinimlerini gidermeleri açısından oldukça faydalı bir yaklaşımdır. Öğretmenlerin sorunun çözümü ile ters düşen bir sonuçla karşılaştığında, bu durumun nasıl ortaya çıktığını sormaları ve oğrenciden durumu acıklamasını istemeleri öğrencide derinliğine düşünme, yaşantı edinme ve tekrar odaklanmayı sağlayabilir. Matematiksel düşünme, yoğun bilişsel etkinlik gerektiren bir durumdur. Bilişsel öğrenme sürecinde, verilen dönütlerin, öğrencide bilişsel çelişkiyi ortadan kaldırması gerekir. Birçok sorunun çözümünde öğretmenin birden fazla yöntem ve tekniği kullanması, öğrenciden de farklı teknikleri işe koşarak çözüme ulaşmasını istemesi, öğrencinin derinliğine düşünmesi ve öğrenmesi açısından bir gerekliliktir. Türkdoğan ve Baki'nin(2012) de öğretmelerin yanlışlara verdikleri dönüte ilişkin tespitleri ile araştırmadan elde edilen sonuçlar örtüşmektedir. Öğretmenlerin bu yönde verdikleri dönütler oldukça anlamlıdır.

Dönütte kapsam teması içinde öğretmenler, konuyu açıklama, gösterme ve çözdürme, problem üzerinde düşündürme, problem çözümünde benzerliklerden ve zıtlıktan faydalanma, ayrıntıları vurgulama ve her öğrenciye yeteri kadar zaman ayırmanın gerekliliği üzerinde durmaktadırlar. Öğretmenlerin verdikleri yanıtlardan dönütün kavramsal olarak kapsamı konusunda zihinlerinin karışık olduğu anlaşılmaktdır. Kimi öğretmenler dönütü, sordukları sorunun çözümlenmiş cevabı; kimileri, öğrencinin öğrenmesini kolaylaştıracak yol gösterici işaret ya da sadece beden dili olarak algılamaktadır.

Öğrenci yanlış yaptığında ona daha onceden öğrendiği matematiksel kavramları ve işlemleri hatırlatmak, çözümlemlerde yeni çıkarımlarda bulunmasını sağlayacak uyaranlar sunmak, etkili bir dönüt verme yolu olarak görülebilir. Esas olan öğretmenlerin, öğrencilerin bazı konularda yanlış yaptıklarını hissettirmesi veya sezdirmesi en etkili dönüt verme şeklidir. Çünkün sezgiye dayalı öğrenme, içsel denetimi ve motivasyonu körükleyici bir özelliğe sahiptir.

Benzer ozelliklere sahip kavramlar, işlemler ve sorunların, daha önce işlenen konularla, sonuçlarla ve çıkarımlarla iliskilendirilmesi, bazı öğrenciler için etkili bir dönüt verme tekniği olabilir. Bu sürecin öğrencinin imajı ile desteklenmesi gerekir. Bu durumda öğretmen, öğrencinin durumu içselleştirmesini sağlayacak şekilde sorunu öğrencinin zihinde canlandırmasını sağlayabilir. Öğrenci konuyu anladığını belirtmiş olmasına rağmen, yanlış yapmaya devam ediyorsa öğretmenin, sorduğu soruyu basitleştirerek, aşamalı çözüm yolunu deneyerek, basit soru çözümünden karmaşık soru çözümüne aşamalı geçişler yaparak çözüm arayabilir. Karmaşık sorunların çözümünde, sorunu küçük anlamlı parcalara bölerek öğretme yolunu deneyebilir. Esas olan öğrencinin, soruyu kendi kendine çözebilmesidir. Doğru yapmaları için öğrencilerin kendi hızlarında ilerlemesini destekleyici uyaranlar verilirken; Yanlış yapmaya devam eden öğrenciler içinde anlaşılması daha kolay olan benzer soruların çözümü gösterilerek işlemin mantığını anlaması, doğru ilişkilendirmeler yapması sağlanabilir.

Dönüte ilkeli davranma teması kapsamında öğretmenlerin görüşleri, öğrencilerde var olan yeterlikten hareket etme, basit örnekler verme, ben dili kullanma, öğrenciye yanlışını gösterme, öğrenciyi motive etmede mevcut doğrularını kullanma ve öğrencilerle bireysel iletişim kurma ve iletişimde zamanlamayı iyi yapma noktasında toplanmaktadır. Öğrencilere verilen her tepkinin bir ölçüsünün ve tutarlılığının olması gerekir. Bu ölçü ve tutarlılıkta öğrencilerin bireysel farklılıklarınıda hesaba katması gerekir. Toplu öğretim sistemi içinde öğrencinin birisine olduğundan fazla açıklayıcı dönüt verilirken bir diğerine verilen dönüt geçiştirilen veya çok sınırlı bir yanıta dönüşmemelidir.

Matematik öğretim sürecinde, öğretmen konunun anlaşılan kısmından başlayarak, öğretime devam etmesi, oğrencinin anladıkları ile yeni öğreneceklerini bütünleştirmesine dayalı stratejiler izlemesi beklenir. Oğretmenin, öğrencinin cevabı bildiğini ama ifade ederken yanlış yaptığını düşünmesi ve bu amaçla tekrarlatması, oğrencinin verdiği cevabı açıklaması için yeteri kadar beklemesi gerekir. Matematik öğretimi sabretme ve öğrencinin zihinsel becerilerini eyleme yönlendirme dersidir. Aksine bazı durumlarda dönüt vermeye zaman kalmaması, öğretimin planlanamayışı, içerik yoğunluğu ve öğrencilerin hazırbulnuşluk düzeyindeki yetersizlikle birleştiğinde öğretmenin üstesinden gelemeyeceği devasa bir soruna dönüşebilmektedir.

Dönüt vermede yaşanan güçlükler temasında öğretmenlerin, öğrencilerde önbilgi eksiği, öğrenci seviyesine inememe, öğrenciye yeteri kadar zaman ayıramama, algılama, anlama ve çözümde zorlanma, aşırı hareketlilik ve öğrenmede direnç, içerik yoğunluğu noktalarında odaklandıkları görülmektedir. Dönütte yaşanan güçlükler teması bağlamında öğretim sürecinde öğretmen, öğrenci ve konu ve sistem kaynaklı ciddi sorunların yaşandığını ortaya koymaktadır.

Matematik dersinin soyut olması, anlatılmak istenen konuların yeteri kadar somutlaştırılamaması öğretmenleri, açık, anlaşılır, kapsamlı ve açıklayıcı dönüt verme

noktasında zorlamaktadır. Matematik tamamen semboller üzerine kurulan kavramsal olarak anlam örüntüsü gerektiren ve bibiriyle sıkı ilişkili işlemler içeren bir alan olduğundan, öğretmenlerin öncelikli olarak konuların öğretimine başlarken sembollere temel teşkil eden matematiksel kavramları anlamlı kılmak için çaba harcamaları gerekir.

Çalışmada elde edilen bulgular, öğretmenlerin dönütün işlevine yönelik hassasiyetlerinin yüksek ve aynı zamanda olumlu yönde olduğunu göstermektedir. Ancak öğrencinin, hazırbulunuşluk düzeyi, geçmiş yaşantıları, akranları ile sınıf içi etkileşim biçimi, dayanışmadan çok yarışmacı bir öğrenme çevresinin öğrenciyi motive etmedeki yetersizliği ve güçlüğü artıran etkenler olarak gözükmektedir. Bu sürecin, öğrenciye yeterli ve kapsamlı dönüt verme, dönüt verirken öğrencinin performansını esas alma, dönütte ilkeli davranma noktalarında öğretmenin işini güçleştirmektedir. Sınıfların kalabalık olması, çoklu etkileşime uygun fiziki donanımın olmaması, matematik içeriğinin yoğun olması, içeriğin müfredata uygun hale getirilememesi, müfredatın yetiştirilme baskısı ve matematik derslerine ayrılan sürenin azlığı gibi husular öğrencilere kapsamlı dönüt vermeyi kısıtlamaktadır.

Öğretmenin rehber olduğu ancak öğrencilerin kendilerini ifade edemediği ortamlarda öğrencilerde ortaya çıkan öğrenme eksikliklerini tespit etmek güçleşmektedir. Kavramsal bilgi eksiği olan öğrencilere sözlü açıklamalar yapmak, işlem kurallarını hatırlatmak tek başına, etkili bir dönüte dönüşmemektedir. İşlem bilgisi eksiği olan öğrenciler için, sembollerin doğru kullanımı veya kavramın günlük yaşamda nerede ve nasıl kullanıldığına yönelik açıklamalar, öğrencinin bu tür eksiğini tamamlamasına yetmemektedir. Öğretmenler, verecekleri dönütleri öğrencilerin gereksinim duyduğu bilgi türüne, işlem şekline, zamana, somut yaşantı eksenine göre öğretim sürecinde birbirini tamamlayan bir bütünlük içinde sunmaları gerekir. Başarılı bir öğretim süreci yakalamak isteyen matematik öğretmeninin dönüt vermeden önce öğrencilerini hazır bulunuşluk düzeyi, motive edebilme, kişilik özellikleri, beklentileri, alışkanlıkları ve tutumları açısından tanıması gerekir. Eksiğinin farkında olmayan bir öğrencinin eksiğini tamamlamasını beklemek mümkün değildir. Öğrencilerin, öğrenmedeki eksiklerini, yanlışlarını ve yetersizliklerini farkına varmaları için, öğretmenlerin öğretme-öğrenme sürecinde verdikleri dönütler, öğrencilerde öz-farkındalık yaratmaya hizmet etmelidir. Bunun için dönütün kapsamlı, belli ilkelere dayalı, etkileşimi artıracak tarzda, verildiğinde öğretim sürecine katkı sağlayabilir. Bu bağlamda matematik öğretmenleri,

- Öğrencilerinin özelliklerini çok yönlü ele alıp, bireysel farklılıklarına uygun dönüt verme yollarını deneyebilirler.

Öğretim sürecini tek düzelikten kurtarmak için, tez-antitez, kuraldan örneğe- örnekten kurala gitme, tümden gelim- tümevarım gibi düşünme süreçlerini zenginleştiren yöntemleri kullanarak, öğrencilerin doğru davranışlarını pekiştirirken, ortaya çıkan yanlışları düzeltmek için yanlışları bir fırsat olarak kullanarak öğrencilerin yanlışlarını fark ettiren hareketle düzeltici dönütler verebilirler.

- Öğretmenler öğrencilerine öğretilen her yeni konuyla ilgili öncelikli olarak konuya temel teşkil eden matematiksel kavramları anlayıp, içselleştirmelerini sağlayıcı somut örnekler verebilirler.
- Öğretmenler, öğrencilerinin öğrenme sürecini yakından takip ederek, hangi noktalarda ne tür eksikliklerinin olduğunu belirleyip, eksiklerini tamamlayıcı dönütler verebilirler.
- Öğrencilerin daha etkin katıldığı ve fikirlerini daha iyi ifade edebildiği öğrenme ortamlarında pekiştirici, destekleyici, yönlendirici dönütler kullanabilirler.
- Öğretmenler için öğrencinin etkin olduğu ve fikirlerini açıkladığı bir öğrenme ortamında öğrencinin zihninden geçenleri belirlemek ve buna uygun dönütler vermek daha kolay olabilir.

Öğretmenler, öğretimin başında öğrencilerinin matematikle ilgili kavram yanılgılarını belirleyip, çözümü ve sonucu control ettikten sonra öğrencilerin eksiklerini ve yanlışlarını ortadan kaldıracak dönütler verme yolunu deneyebilirler.

Teşekkür

Bu çalışma, Ahi Evran Üniverstesi BAP birimi tarafından EGT. A3.16.15 nolu proje ile desteklenmiştir.

5. KAYNAKÇA

Akyol, H. (2007). *Vygotsky, Piaget ve Yapılandırmacı Okuma Eğitimi*. VI. Ulusal Sınıf Öğretmenliği Kongresi Bildiri Kitabı, Eskişehir.

Black, P. & William, D. (1998). Assessment and classroom learning. *Assessment Education,* 5 (1), pp.7-74.

Brookhart, S.M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 3*(1), pp. 3-12.

Clark, K. & Dwyer, F. M. (1998). Effect of differenttypes of computer-assisted feedbacks trategies on achievement and response confidence. *International Journal of Instructional Media, 25*(1), pp.55-63.

Carvalhoa, C. & Santosa J. & Conboya, J. & Martinsa D. (2014). Teachers´ feedback: Exploring differences in students´ perceptions. *Procedia - Social and Behavioral Sciences* 159, pp. 169-173.

Crooks, T.J. (1988). The impact of classroom evaluation on students. *Review of Educational Research,* 5, pp. 438-481.

Dempsey J.V., Litchfield B.C. & Driscoll M.P., (1993). Feedback, Retention, Discrimination Error, and Feedback Study Time, Journal of Research on Computing in Education, 25: 3, pp. 303-326.

Duhon, G., House, S., Hastings, K., Poncy, B. & Solomon, B. (2015). Adding immediate feedback to explicit timing: An option for enhancing treatment ıntensity to ımprove mathematics fluency. *Journal of Behavioral Education*, *24*(1), pp. 74-87.

Eraz, G. & Öksüz, C. (2015). Sınıf Öğretmenlerinin Öğrencilerin Ders Dışı Matematik Etkinliklerine Ilişkin Uyguladıkları Geribildirimlerin Etkisi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 36*, ss.105-119.

Erişen, Y. (1997). Öğretim Elemanlarının Dönüt ve Düzeltme Davranışlarını Yerine Getirme Dereceleri. *Kuram ve Uygulamada Egitim Yönetimi Dergisi, 3*(1), ss. 45-62.

Foote, C.J. (1999). Attribution feedback in the elementary classroom. *Journal of Research in Childhood Education*, 13(2), 155-166.

Hattie, J. & Timperley, H. (2007). The power of feedback, *Review of Educational Research*, 77 (1), pp. 81-112.

Kluger, A.N. & Denisi, A. (1996). The Effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 2(2), pp. 254-284.

Köğce, D. & Baki, A. (2012). İlköğretim Matematik Öğretmenlerinin Geribildirim Kavramına İlişkin Inanışları, X. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi, 27-30 Haziran, Niğde.

Labuhn, A.S., Zimmerman, B.J., & Hasselhorn, M.(2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning, 5*(2), pp. 173-194.

Mory, E.H. (2004). *Feedback research revisited*. In D. Jonassen, (Ed.), Handbook of Research on Education Communications and Technology (pp. 745-783). Mahwah, NJ: Lawrence Erlbaum Associates.

Manouchehri, A. (2007). Inquiry-discourse mathematics instruction. *Mathematics Teacher,* 101 (4), pp. 290-300.

Manouchehri, A. & St. John, D. (2006). From classroom discussions to group discourse. *Mathematics Teacher, 99* (8), pp. 544–551.

Naroth, C. (2010). Constructive teacher feedback for enhancing learner performance in mathematics.

[serial online]. n.d.; Available from: Networked Digital Library of Theses & Dissertations, Ipswich, MA. 21 Ağustos 2016 da ulaşılmıştır.

Nordstrom, C.R., Wendland, D. & Williams, K.B. (1998). "To err is human": An examination of the effectiveness of error management training, *Journal of Business and Pschology,* 12, 3, pp. 269-282.

Kahl, S. (2005). Where in the world are formative tests? Right under your nose! Education Week, 25 (4), 38.

Looney, J. (Ed.). (2005). *Formative assessment: Improving learning in secondary classrooms*. Paris, France: Organisation for Economic Cooperation and Development.

Peker, R. (1992). Geri Bildirimin Üniversite Öğrencilerinin Ölçme Ve Değerlendirme Dersindeki Başarısına Etkisi. *UludağÜniversitesi Eğitim Fakültesi Dergisi*, 7(1), ss. 31-39.

Rakoczy, K., Klieme, E., Bürgermeister, A. & Harks, B. (2008). The interplay between student evaluation and instruction. grading and feedback in mathematics classrooms. *Zeitschrift für Psychologie, 216*, pp. 110–123.

Rakoczy, K., Harks, B., Klieme, E., Blum, W. & Hochweber, J. (2013). Written feedback in mathematics: Mediated by students´ perception, moderated by goal orientation. *Learning and Instruction*, 27, pp. 63-73.

Roschelle, J., Rafanan, K., Bhanot, R., Estrella, G., Penuel, B., Nussbaum, M., & Claro, S. (2010). Scaffolding group explanation and feedback with handheld technology: Impact on students' mathematics learning. *Educational Technology Research and Development*, (4), pp. 399-404.

Sadler, D.R. (1998). Formative assessment: revisiting the territory. *Assessment in Education, 5*(1), pp. 77-84.

Saracaloğlu, A.S., Gencel, İ.E. & Çengel, M. (2011). Öğrenci ve Öğretmen Görüşleri Açısından Lise Öğretmenlerinin Öğretme Sürecindeki Yeterlikleri, *Adnan Menderes Üniversitesi Eğitim Fakültesi Eğitim Bilimleri Dergisi, Aralık, 2 (2), ss.77-99.*

Stevenson, C.E., Heiser, W. J. & Resing, W. C. M. (2013). Working memory as a moderator of training and transfer of analogical reasoning in children. *Contemporary Educational Psychology*, 38(3), pp.159-169.

Stone, N.J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*, pp. 437-475.

Şahin, M. (2015). Öğrenme ve Öğretme Sürecinde Uygulanan Dönüt Etkinliği ile İlgili Öğretmen Adaylarının Görüşlerinin İncelenmesi, *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 15(1), ss.*247-264.

Santagata, R. (2002). *When student make mistake: Socialization practices in ıtaly and the united states*, Doctoral Dissertation, Los Angelels: University of California, Philosophy in Psychology.

Turkdoğan, A. Baki, A. (2012). İlköğretim İkinci Kademe Matematik Öğretmenlerinin Yanlışlara Dönüt Vermede Kullandıkları Donüt Teknikleri, *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 45*, (2), ss.157-182.

Warden, C.A. (2000). EFL business writing behaviors in differing feedback environments. *Language Learning*, 50 (4), pp. 573–616.

Wigfield, A., Klauda, S.L., & Cambria, J. (2008). Influences on the development of academic self-regulatory processes. In B.J. Zimmerman, & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp.33-48). New York: Routledge.

Zimmerman, B.J., & Martinez-Pons, M. (1992). Perceptions of efficacy and strategy use in the self-regulation of learning. In D. H. Schuck & J. L. Meece (Eds.), *Student perceptions in the classroom*. Hillsdale, NJ: Lawrence Erlbaum.

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Investigation of Equating Error in Tests with Differential Item Functioning

Meltem Yurtçu, Cem Oktay Güzeller

To cite this article: Yurtçu, M. & Güzeller, C.O. (2018). Investigation of Equating Error in Tests with Differential Item Functioning. *International Journal of Assessment Tools in Education*, 5(1), 50-57. DOI: 10.21449/ijate.316420

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 50-57

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Investigation of Equating Error in Tests with Differential Item Functioning

Meltem Yurtçu[*1], **Cem Oktay Güzeller**[2]

[1]Hacettepe University, Faculty of Education, Department of Measurement and Evaluation in Education, Turkey
[2]Akdeniz University, Faculty of Tourism, Turkey

Abstract: In this study purposes to indicate the effect of the number of DIF items and the distribution of DIF items in these forms, which be equalized on equating error. Mean-mean, mean-standard deviation, Haebara and Stocking-Lord Methods used in common item design equal groups as equalization methods. The study included six different simulation conditions. The conditions were compared according to the number of DIF items and the distribution of DIF items on tests. The results illustrated that adding DIF items to tests were equated caused an increase in the errors obtained by equating methods. We may state that the change in errors is lowest in characteristic curve transformation methods, largest in moment methods depending on the situations in these conditions.

ARTICLE HISTORY

Received: 27 May 2017
Revised: 13 June 2017
Accepted: 15 September 2017

KEYWORDS

Equation error,
DIF,
IRT equation methods,
Large Scale Exams

1. INTRODUCTION

Countries participate in large-scale tests at international or national level or prepare and implement large-scale examinations in order to evaluate the educational systems or to place students in upper level educational institutions. These implemented tests are prepared in various forms in order to ensure reliability and to be able to compare the test scores of individuals taking these tests at different times. It is necessary to equate their scores in order to be able to make a comparison of scores of people taking these test forms or to make a comparison of the difficulty of exams prepared for the same purpose (Dorans & Holland, 2000; Dorans, 2004; Kim, Walker & McHale, 2010).

Through procedures applied to the scores obtained from the test forms measuring the same construct, it is possible to make these scores interchangeable regardless of when and to whom these test forms are applied (Kolen & Brennan, 2004; Dorans & Holland, 2000). Test equating is a statistical and psychometric technique used for the adjustment of scores from different tests measuring the same construct in order to compare scores obtained from various forms of that test (Dorans & Holland, 2000; Skaggs, 2005). Felan (2002) points out that the scores obtained from different tests can be placed on a single scale and compared simultaneously via the statistical relationship established between the scores obtained from two

*Corresponding Author E-mail: meltem.yurtcu@gmail.com cguzeller@gmail.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.316420**

different forms measuring the same construct. According to a definition by Angoff (1971), test equating is the process of converting the unit scale of a test form to the unit scale of another test form. Kim and Hanson (2002) express equating as interchangeability of test forms after procedures applied to points from these test forms. In principle, the process of establishing the relationship between raw or scaled points used in two or more test forms is described as equating (Skaggs & Lissitz, 1986). The conditions required to be able to do equating are measuring the same construct, having equal reliability, equity, and invariance between groups (Dorans & Holland, 2000; Lord, 1980; Swaminathan & Gifford, 1983).

The right decision making end of these large scale exams that are extremely important for societies depends on reliability and validity of exams. Especially in equating of large-scale, there are a lot of situation that threaten reliability and validity. The some of the situations stem from multiple sources including measurement error, sampling error, measurement disturbances and administrative challenges. Measurement error usually refers to inaccurate associated with a measuring instrument (Wu, 2010). Depending on the equating method and pattern, the error emerging as a result of equating is of two types: random and systematic (Kolen, 1988; Felan, 2002). While random error that stems from answerer sampling is defined as standard error of equating (Kolen & Brennan, 2004); the other type of equating error, which is also known as equating bias, stems from violation of axioms or from biasedness (Zeng, 1991). Biasedness arises as a result of evaluation of an item with differential item functioning (DIF) by specialist opinion and involves sensitivity and differential item functioning analysis (Hambleton, 2006; Sireci & Mullane, 1994; ETS, 2009).

DIF surfaces as individuals with similar ability level but are in different subgroups differ in their probability for answering test items (Osterlind, 1983; Zumbo, 1999). Differential item function is of two types: uniform and non-uniform. It is considered uniform if the probability an item being answered correctly contains DIF in favor of a specific group for all ability levels but non-uniform if it contains DIF in favor of different groups at different ability levels (Zumbo, 1999). Investigation of differential item functioning (DIF) is with outmost important on the accuracy of the decisions taken as a result of large-scale examinations for societies when comparing measures across different groups (Lai, Teresi & Gershon, 2005; Swaminathan & Rogers, 1990). The presence of a DIF item(s) in the test, an indication of bias, will cause the obtained scores to be misleading (Zieky, 2002; Osterlind, 1983).

In the context of this study, the aim is to investigate the effect on the equating error obtained from the IRT-based equating methods according to the test containing DIF items and the number of DIF items in two tests with the same item parameters during the process of placing the points obtained from these tests on the same scale. Equalization of tests containing DIF items with item response models takes place in the literature using different methods and conditions (Demirus, 2015; Huggins, 2014). However, differentiation of the number of DIF items and the distribution of DIF items in test forms which be equalized in common item design equal groups makes this work unique from other studies. In this respect it will be contribute to literature. In this line, the basic research question may be formulated as:

“What are the effects of the number of DIF items in tests and of the tests containing DIF items on the equating error during the process of placing two math tests measuring the same construct on the same scale?”

2. METHOD

In this study purposes to indicate the effect of the number of DIF items and the distribution of DIF items in these forms, which be equalized on equating error. This is a basic research study in essence since it investigates the effect of the number of DIF items present in forms on

equating error with respect to the forms including DIF items by using IRT equating methods on common item pattern in equal groups.

2.1. Data Collection

Here, the study was conducted on the data set generated from the 2013-2014 TEOG exam on the basis of the assumption that the tests were taken by individuals with equal ability. Two different math test forms were generated with Wingen2 program by using item parameters in the math test of this exam. These forms are comprised of a medium-length test containing 15 common items aside from a set of 40 parallel questions. Hence, scores obtained from two tests containing 55 items per each were on the same scale. The item parameters of the math test were 0.20-0.76 for parameter a, 0.34-0.83 for parameter b, and 0.25-0.40 for parameter c. The common item pattern in equal groups was used as a pattern in equating. The forms A and B with 40 items per each were generated for different conditions in accordance with the three-parameter logistic model scored as 1-0 regarding the Item Response Theory models. Since the common form was so as to reflect A and B tests, it was generated by using the same parameters. The forms were generated to measure the same construct unidimensionally. For the ability distribution of the groups taking these forms, 1000 answers with normal distribution were generated so as the mean is 0 and standard deviation is 1. There are items with uniform DIF at B (medium) level in the common test and in the basic test on the generated forms. The DIF items were obtained as in favor of single group (in favor of males in TEOG); sizes of focus and reference groups are equal.

In order to answer the research question, six different conditions were considered: two different situations for number of DIF items (5 and 10) and three different situations for the test form containing the DIF items (form A, form B, and the Common form). The patterns of conditions are given in Table 1.

Table 1. The conditions determined with respect to the number of DIF items on forms and on the forms containing DIF items.

	Number of Items	**Total of 5 DIF Items**			**Total of 10 DIF Items**		
Form A	40	5 DIF Items	3 DIF Items	-	10 DIF Items	5 DIF Items	-
Form B	40	-	-	-	-	-	-
Common Form	15	-	2 DIF Items	5 DIF Items	-	5 DIF Items	10 DIF Items
		Condition 1	**Condition 2**	**Condition 3**	**Condition 4**	**Condition 5**	**Condition 6**

As it is seen in Table 1, six different conditions were obtained on the basis of different number of DIF items contained and the test forms these DIF items were on after forms A and B were generated as basis. Attention was paid to not to place the DIF items on tests consecutively.

2.2. Data Analysis

The common form was included in scores as internal anchor test in the study. Since the data belonging to test forms used in this study display similar difficulty and selectivity means, horizontal equating was done among these test form. The same parameters were used for common form data.

Separate conjecture methods were used for equating pattern used. PARSCALE 4.1 program was used for conjecture of parameters, IRTEQ program was used for test equating and

scaling. Data derivation and equating process were repeated 25 times for each condition and each method.

The root mean square deviation (RMSD) value was used in equating the test scores that the individuals with same ability level have received from different test forms. The RMSD values obtained from Mean-Mean, Mean-sigma, Stocking-Lord, Heabara equating methods were obtained by averaging 25 repeats.

3. FINDINGS

The six conditions were considered for the comparison of the equating error obtained by different IRT equating methods on the basis of the number and distribution of DIF items. In order to compare the condition as criteria, the equating errors in condition where both test forms do not contain DIF items.

Firstly, the condition where the 7^{th}, 12^{th}, 23^{rd}, 26^{th}, and 37^{th} items in the first 40 questions of the basic test, which is called test A and is among the math test to be equated, display uniform DIF with a difference of 0.6 at B level and there is no DIF item in the first 40 questions of the common test and form B was considered. This condition where there are five DIF items in the basic test and no DIF items in common test and form B is called Condition 1.

Condition 2 was created where DIF items are present both in the common test and the basic test, as number of DIF items is kept same. Under this condition, it is assumed that there are three DIF items, the 5^{th}, 17^{th}, and 33^{rd} items, in the first 40 questions of the basic test; and there is DIF in the 47^{th} and 53^{rd} items of the common test.

Condition 3 was created to analyze the RMSD value where DIF items are present only in the common test, as number of DIF items is fixed. Under this condition, it is assumed that there is DIF in the 51^{st}, 52^{nd}, 53^{rd}, 54^{th}, and 55^{th} items only in the common test form of the math test.

In order to investigate the effect of the change in the number of DIF items on equating error, the number of DIF items in the first 40 questions of the basic test is considered to be ten. Items that were considered as having DIF are the 5^{th}, 7^{th}, 12^{th}, 17^{th}, 23^{rd}, 26^{th}, 29^{th}, 33^{rd}, 37^{th}, and 40^{th} items. The condition where there is no DIF item in the first 40 questions of the common test and form B is called Condition 4.

Condition 5 was created which tests the DIF items are present in while the number of DIF items in tests to be equated is taken as ten and the number of DIF items is fixed. For this condition, it is assumed that the 7^{th}, 12^{th}, 23^{rd}, 26^{th}, and 37^{th} items of the first 40 questions on A test and the 51^{st}, 52^{nd}, 53^{rd}, 54^{th}, and 55^{th} items of the common test have DIF.

Created condition 6 where there are ten DIF items only in the common test is assumed that only the 46^{th}, 47^{th}, 48^{th}, 49^{th}, 50^{th}, 51^{st}, 52^{nd}, 53^{rd}, 54^{th}, and 55^{th} items on the common test form have DIF.

We examined RMSD equating errors of equating done by four methods for 6 conditions and math test forms without DIF as scaling method. The equating errors, which were obtained as the points taken from tests A and B belonging to these conditions were placed on same scale, were investigated with respect to IRT equating methods. These values were shown in Table 2.

Table 2. The RMSD equating errors of equating done by four methods for conditions where math test forms without DIF.

	Mean- Mean	Mean-Sigma	Haebara (HB)	Stocking-Lord (S-L)
The equating errors for test forms without DIF	0.057616	0.179619	0.17014	0.171374
Condition 1	1.14101	0.842776	0.98555	0.597466
Condition 2	0.348804	0.511489	0.328713	0.295562
Condition 3	0.39065	0.588079	0.308391	0.291414
Condition 4	1.165186	0.886565	0.600028	0.606109
Condition 5	0.646586	0.915705	0.546247	0.519187
Condition 6	0.318883	0.69995	0.352803	0.332708

condition 1: five DIF items in the test A and no DIF items in common test and form B
condition 2: five DIF items in the test A and two DIF items in test B
condition 3: five DIF items in the common test of the math forms
condition 4: ten DIF items in test A and there is no DIF in the common test and form B
condition 5: ten DIF items in test A and five DIF items test B
Condition 6: ten DIF items in the common test of the two math forms

When the tests forms don't include DIF items, the lowest error among the IRT equating methods looks to be with Mean-Mean method. It is followed by the equating error calculated by the Haebara method. The highest error was obtained by Mean-sigma method.

In condition 1, the lowest error among the IRT equating methods looks to be with Stocking-Lord method in conditions B. It is followed by the equating error calculated by the Mean-sigma method. The highest error was obtained by Mean-Mean method.

In condition 2, condition 3 and condition 5 the lowest error among the IRT equating methods looks to be given by Stocking-Lord method. It is followed by the equating error calculated by the Heabara method, one of the characteristic curve methods. The highest error was produced by Mean-sigma method in this condition.

When condition 4 is examined, the lowest error among the IRT equating methods looks to be given by Haebara method. Following this method, the points obtained by the Stocking-Lord method look to have the next lowest error. It is observed that the highest error was obtained by Mean-sigma method.

When Condition 6 is examined, the lowest error among the IRT equating methods looks to be given by Mean-Mean method. The error coefficient obtained by the Heabara method follows. It is observed that the highest error was obtained by Mean-sigma method.

4. RESULTS AND CONCLUSION

Changes in the curriculum, such as test structure, test length, and retention exposure can create bias among individuals (Stocking & Lewis, 1998). The presence of questions, which may create a bias in favor of a specific group in one or two of the tests being equated, will affect the validity of this test (Osterlind, 1983; Zieky, 2002). It is also important to test whether the anchors items included in the test have DIF (Klein & Jarjoura, 1985; Cook & Petersen, 1987).

In accordance with the purpose of the study, it was investigated that inclusion of the DIF items in test equating process casts doubt on the accuracy of the scores generated as a result of equating. RMSD was used as the criteria value because of providing an estimate by combining the random and systematic equating error (Puhan, 2010; Sinharay & Holland, 2007) and these RMSD values of IRT equating methods were considered were compared to each other. Variations in the RMSD value, which was considered as the equating error, were examined with

respect to the number of DIF items and with respect to which test forms have the DIF items among the tests to be equated.

Presence of DIF items in any of tests to be equated causes a decrease in errors calculated by all IRT equation methods. While increasing the number of DIF items only in test A causes an increase in errors for all methods except for Haebara method, increasing the number of the DIF item only in common test causes increase in errors for all methods except for the mean- mean method. Increase in the number of DIF items both in the common test and the basic test causes an increase in error calculated by all methods. When conditions that include the same number of DIF items in common test are compared, the presence of DIF items in the basic test also increases the error.

That there are DIF items in both tests causes it to have less error than the condition where only test A has DIF items except for mean-sigma method in competing condition 5 and condition 4. To see this, it can be compared to condition 1 and condition 2; condition 1 and condition 3; condition 4 and condition 6.

When it is examined all conditions including condition where both test forms do not contain DIF items, generally it can be seen that lowest equation errors are obtained by Stocking-Lord method and the highest error was obtained by Mean-sigma method during equating done in the study.

According to research studies that have a common finding is that item characteristic curve methods give more accurate than moment methods (Beguin, 2002; Kim & Cohen, 1992; Way & Tang, 1991; Stocking & Lord, 1983; Ogasawara, 2001). Kilmen and Demirtaşlı (2012) also express their study that equation errors are obtained by Stocking-Lord method indicate less errors than other IRT methods. The c parameter is never considered in the calculation of the scale factor since the mean-sigma and mean-mean methods derive the scaling factors from the descriptive statistics of the distribution of b-parameters. We may state that the equating error obtained by Mean-Mean and mean-sigma methods is higher due to added DIF items being uniform and being a result of a change of 0.6 unit at B level.

In the literature, there is very little work that compares the methods of equalization on this subject. Demirus (2015), who examines the effects of items with DIF on the real data, in case the anchor items display uniform DIF for a group, the mean-mean method produces the largest error, the mean-sigma method yields the smallest. On the anchor items without DIF the biggest equating error has been obtained by mean-sigma method and smallest equating error has been obtained by Stocking-Lord and Haebara methods. This is partly similar to our findings.

In future studies, the status of mixed-structure test that includes DIF items can be examined. The DIF level taken the uniformly in this study can be considered at many different levels. In addition, as a different dimension of this study, it is possible to examine how the results will be observed when the skill levels of the groups receiving the tests to be equal are different.

5. REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 508-600). Washington, D.C.: American Council on Education.

Béguin, A. A., Hanson, B. A. & Glas, C. A. W. (2000). Effect of unidimensionality on separate and concurrent estimation in IRT equating. Paper presented at the *Annual Meeting of the National Council on Measurement in Education*, New Orleans, LA. Available from http://www.bah.com/papers/paper0002.html

Cook, L. L. & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225–244

Demirus, K. B. (2015). *Ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunda test eşitlemeye etkisinin farklı yöntemlerle incelenmesi.* Doktora Tezi, Ankara: Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü.

Dorans, N. J. & Holland, P. W. (2000). Population invariance and the equatability of tests: basic theory and the linear case. *Journal of Educational Measurement, 37* (4), 281- 306.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43-68.

Educational Testing Service. *Guidelines for fairness review of assessment*. Retrieved May 22, 2015 from http://www.ets.org/Media/About_ETS/pdf/overview.pdf

Felan, G. D. (2002). Test Equating: Mean, Linear, Equipercentile and Item Response Theory. Paper presented at the *Annual Meeting of the South West Educational Research Association*, Austin.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care. 44*(11), 182-188.

Huggins, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement. 74*(4), 627-658.

Kilmen, S. & Demirtaslı, N (2012). Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia - Social and Behavioral Sciences, 46*, 130-134.

Kim, S. & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement*, *26*(3), 255-270.

Kim, S. & Cohen, A.S. (1992). Effects of linking methods on detection of DIF. *Applied Psychological Measurement*, *29*(1), 51-56.

Kim, S., Walker, M.E. & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47* (1), 36-53.

Klein, L. W. & Jarjoura, D. (1985). The importance of content representation for common item equating with non-random groups. *Journal of Educational Measurement, 22*, 197-206.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement Issues and Practice, 7* (4), 29-36.

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking* (2nd edition). USA: Springer.

Lai, J. S., Teresi, J. & Gerson, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes, *Evaluation & The Health Professions, 28*(3), 283-294.

Lord, M. F. (1980). *Application of Item Response Theory to Practical Testing Problems.* New Jersey: Lawrence Erlbaum Associates Publishers.

Ogasawara, H. (2001). Item response theory true score equating and their standard errors. *Journal of Educational Behavioral Statistics*, *26*(1), 31-50.

Osterlind, J. S. (1983). *Test item bias*. London Sage Publications.

Puhan, G. (2010). A comparison of chained linear and post stratification linear equating under different testing conditions. *Journal of Educational Measurement*, *47*(1), 54–75.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement, 27*(4). 361-370.

Sinharay, S. & Holland, P.W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249–275.

Sireci, S.G. & Mullane, L. A. (1994). Evaluating test fairness in licensure testing: The sensitivity review process. *CLEAR Exam Review*. *5*(2), 22-27.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42* (4), 309–330.

Stocking, M.L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed- and true- score equating procedures*. (ETS Research Report NO. RR-88-41). Princeton, NJ: Educational Testing Service.

Stocking, M.L. & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Swaminathan, H. & Gifford, J.A. (1983). Estimation of parameters in the three parameter latent trait model. In D. Weiss (Ed.), *New horizons in testing.* New York: Academic Press.

Zeng, L. (1991). Standard errors of linear equating for the single-group design (ACT Research Report 91-4). Iowa City, IA: American College Testing.

Zieky, M. (2002). Ensuring the fairness of Licensing Tests. *CLEAR Exam Review*. *12*(1), 20-26.

Zumbo, B.D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Way, W. D. & Tang, K.L. (1991). A comparison of four logistic model equating methods. Paper presented at the *Annual Meeting of the American Educational Research Association*, Chicago.

Wu, M. (2010). Measurement, Sampling, and Equating Errors in Large-Scale Assessments. *Educational Measurement: Issues and Practice, 29* (4), 15–27.

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Comparing Physics Textbooks in Terms of Assessment and Evaluation Tools

Zeynep Başkan Takaoğlu

To cite this article: Başkan Takaoğlu, Z. (2018). Comparing Physics Textbooks in Terms of Assessment and Evaluation Tools. *International Journal of Assessment Tools in Education*, 5(1), 58-72. DOI: 10.21449/ijate.320214

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 58-72

http://www.ijate.net

e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Comparing Physics Textbooks in Terms of Assessment and Evaluation Tools

Zeynep Başkan Takaoğlu[*1]

[1]Gümüşhane University, School of Health, Gümüşhane, Turkey

Abstract: Assessment and evaluation instruments provide teachers the opportunity of shaping education in the beginning, contributing to education during the process and evaluating education at the end of the process. Textbooks, on the other hand, are resources that present the aforementioned contributions to teachers at first hand. Thus, the study aims to compare the distribution of assessment and evaluation instruments in the physics textbooks being used in the academic year of 2011- 2012 and 2016-2017 according to units, settlement within units and types of assessment instruments that are used. For that purpose, 9, 10, 11 and 12th grade textbooks being used in physics lessons in the academic year of 2011-2012 and 2016-2017 were examined via document analysis method. As a result of the study, it was determined that the highest number of assessment instruments in physics textbooks from two different years was encountered in the unit of force and motion. The reason for this unit having higher number of questions could be associated with higher number of mathematical operations in the unit intended for allowing students to overcome their mathematical deficiencies by practicing such questions. It was observed that the number of questions was increased especially in the books being used in the academic year of 2016-2017 and alternative assessment instruments were fewer than traditional assessment instruments. Traditional assessment instruments are still used very frequently in the textbooks, which proves the effect of traditional approaches in assessment and evaluation. Another reason for this condition is that a result-oriented evaluation is used in the university entrance exam. In the light of these results, it is suggested to make the university exam student-centered rather than making an arrangement in textbooks.

ARTICLE HISTORY

Received: 09 June 2017

Revised: 21 September 2017

Accepted: 02 October 2017

KEYWORDS

Physics textbooks,
Assessment and evaluation,
Physics education

1. INTRODUCTION

Change has an important place in human life. Individuals feel the need to develop and change themselves, depending on their environment, living conditions and cultural factors since they are born. The education and training activities carried out in the schools are important for the implementation of these changes in the lives of the individuals. Education and training institutions need to constantly renew and develop themselves in order to have the power of competition and sustain their assets, reach their goals effectively and efficiently

[*]Corresponding Author E-mail: zeynepbaskan@hotmail.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.320214**

(Çalık, 2003). The changes in the education programs need to be regulated in order to gain new values together with the developing society and to gain the attitudes, values, and information that are necessary for the changes taking place in relation to culture, politics and the external world (Erdoğan, 2015). For this reason, changes in educational curricula are carried out by taking into consideration learning environments, schools, teachers, students and learning materials (Küçüközer & Bostan, 2007).

A curriculum developed in line with a specified philosophy helps teachers organize teaching and learning activities while writing textbooks, selecting technology and teaching materials to use (Kaya, 2013). Textbooks play an important role in presenting to practitioners by taking changes in the industry, technology, and other fields into account. At the same time, they can be expressed as resources to help to narrate the basis of teaching programs (Yiğit, Alev, Özmen & Akyıldız, 2009). Textbooks avail not only to teachers to lecture systematically, the better use of their power and present the lesson, but also to students to review courses whenever they need and to learn by going over the lessons that are not being understood (Küçükahmet, 2003).

Since 1992, a renewal study has not been carried out in the physics curriculum and the same curriculum and textbooks have been used. However, as a result of the renovation studies carried out in primary school science courses and consequently the constructive approach-based studies practiced in the schools, renewal studies in secondary education, which is the continuation of primary schools, became inevitable and from 2007 onwards the physics curriculum gradually entered into force. It has been taken into consideration that learning experience gets easier, meaningful and permanent in natural environment when needed and that the association with real life events to teach physics concepts and laws in the physics course curriculum (Arslan, Tekbıyık & Ercan, 2012). However, due to various problems while the program is running and the need for renewal with the developing technology, the physics curriculum has been updated in 2013. Textbooks prepared in line with the updated program started to be used gradually starting from the 2013-2014 academic year. Features of the renewed physics curriculum are stated as; the clarification of the classes with accompanying units by Yiğit (2013), the step-by-step application by teachers of the models or methods mentioned in the books, the liberalization of the program structure and the decrease in the number of gains.

One of the innovations seen in physics textbooks with the curriculum renewed in 2007 has been in the part of assessment and evaluation. In addition to the traditional approach, alternative assessment and evaluation are now being used for assessment and evaluation. In the textbooks, process evaluation, authentic tasks, application of information, creation of evaluation criteria with clear and significant criteria, performance tasks and evaluation with multiple methods have come into the forefront. Those contribute to the success of the students, student-centered approach, multidimensional evaluation, evaluation of multiple truths, feedback, continuous assessment, evaluation of senior skills and clear results (Gömleksiz, Yıldırım & Yetkiner, 2011). In the physics curriculum renewed in 2013, following topics are emphasized in the area of assessment and evaluation; "to associate teaching and assessment and evaluation with each other, to make plans for assessment, to prepare valid and reliable assessment tools, to use various assessment methods, to use metrics that require the use of information instead of recall, to measure the learning and development of the learners frequently, to measure not only results but also process, to measure the goals stated in the curriculum, to make use of registration and scoring methods, to make evaluation and feedback at the beginning, at the end of and during the education" (MEB, 2013).

Due to the significant contributions to education and teaching, studies conducted in the field of assessment and evaluation also vary. However, studies are usually focused on opinions

of teachers or teacher candidates (Ataman & Kabapınar, 2012; İzci, Göktaş & Şad 2014; Öztürk, Yalvaç Hastürk & Demir, 2013; Peker & Gülle, 2011; Sağlam Arslan, Avcı & İyiybil, 2008; Tay, 2013). The way in which ideas of teachers change as much as the ideas need to be examined in terms of different variables. In this context, textbooks are the most used resources for the teachers during the course preparation (Nakiboğlu, 2009). For this reason, the examination of textbooks in terms of assessment and evaluation will be accompanied by an evaluation of teachers' opinions. In this regard, one more variable among factors that influence teachers' opinions will come to light, so a different dimension will be added to the work in this direction. Despite the fact that studies on the field of assessment and evaluation in the textbooks are not available for physics courses, they are available in Biology, Turkish, Science and Mathematics courses (Arslan & Özpınar 2009; Çetin & Çakır, 2013; Göçer, 2008; Tabak, 2007; Taşdere, 2010). However, in some of the studies, assessment and evaluation are examined in one section, while others focused on assessment and evaluation-program adaptation. Assessment and evaluation studies carried out for textbooks should be emphasized in terms of physics lecture.

The subjects such as visual evaluation, content-curriculum adaptation were investigated in the studies carried out considering the physics textbooks (Ayvacı & Devecioğlu, 2013; Çepni, Ayvacı, Şenel Çoruhlu & Yamak, 2014; Güzel & Adıbelli, 2011). Research has been carried out in the textbooks examined, focusing on only one class, without considering all levels. In the studies carried out on these books, mostly textbooks which were gradually used in 2007 were taken into consideration. The evaluations were carried out by referring to teachers' or teacher candidates' opinions. Teachers need to be supported by studies that take into consideration direct textbooks because they can initially resist to the implementation of the program and can assess it in this direction. For this reason, studies should be carried out by the researchers to examine the textbooks in line with the criteria determined in the research.

The revised physics curriculum in 2013 and studies on textbooks that have been in use since that date are still very new. In the studies carried out, the focus is mainly on comparing the structure and content of the program and examining the objectives of the program rather than examining textbooks (Göçen & Kabaran, 2013; Eke, 2016; Kotluk & Yayla, 2016; Yiğit, 2013). The comparison of the physics curriculum was carried out by taking into consideration the various items found in the curriculum. In addition, examining the gains in the program in the priority of various models or theories can be given as an example of the work carried out on the program. However, no study has been done on textbooks prepared in accordance with the 2013 curriculum.

As it can be understood from the literature reviewed, physics textbooks have not been adequately examined in terms of assessment and evaluation. The examination of physics textbooks, which are among the most important resources of teachers, in terms of assessment and evaluation is also very important for the renewal and development studies to be carried out in the programs and books. The studies carried out for the assessment and evaluation in the textbooks are an important source for the development of other teaching fields.

The main purpose of the study is to compare how the assessment and evaluation tools in the physics textbooks used in the 2011-2012 and 2016-2017 academic years are distributed according to the types of units and types of measuring instruments. The reason for choosing textbooks used in these years is due to the fact that figural arrangements have been made in the physics curriculum in previous years. Two sub-problem responses were sought in this direction.

1. What is the distribution of assessment and evaluation instruments in the physics textbooks of both years according to the units and the placement in the units?

2. Which assessment and evaluation instruments were included in the Physics textbooks of both years?

2. METHOD

The origin of the study is based on the qualitative research design. Qualitative research takes into consideration the qualitative data collection methods such as observation, interview, document analyses and takes the events and situations as a whole in their natural environment (Yildirim, 1999).

In this study, document analysis in the qualitative research category was used. In this process, the sources and the required information are examined, and then the thoughts and ideas to be reached get clearer with the syntheses made and the classification of the data according to the specific properties (Çepni, 2007). The method of document review is divided into two areas as general screening and content analysis (Karasar, 2007). Content analysis is to analyze the printed and visual materials thematically by specific categories (Saban, 2009). For this reason, in the scope of the document examination in the study, the data were analyzed in accordance with the content analysis.

In this context, the 9th, 10th, 11th and 12th class physics textbooks prepared by the Ministry of National Education in the 2011-2012 academic year and the 9th and 10th class physics textbooks belonging to Tuna Printing Company and the 11th and 12th class textbooks belonging to the Dikey publishing in the 2016-2017 academic year are taken into consideration. In the study, assessment and evaluation tools at the beginning of the units, through the units and at the end of the units with these units are examined and the results are compared.

2.1. Analysis of Document Review Data

In the analysis process of the data, the documents were analyzed using two different criteria for each textbook. In the first phase of the study, classes and units were taken into consideration and a categorization was carried out for questions. In the second stage, the examined textbooks are classified according to the assessment and evaluation tools they contain. In the data analysis process, questions in physics textbooks are classified separately according to their placements as at the beginning, through, and at the end of units. Subsequently, these questions were presented in a single table comparing the different years, taking into consideration of the units. In the second phase of the study, assessment and evaluation tools were categorized according to their types. Expert opinions were consulted at unsteady points and the question was placed in an appropriate category in this direction. After the necessary data were obtained, the assessment and evaluation tools were grouped in itself included in each class were grouped composing first tables. Thus, 8 tables belonging to different classes appeared. In the second stage, these tables were combined taking into account the assessment and evaluation tools. Here, questions at the beginning of, through, and at the end of units for all classes are shown comparatively.

In the study of physics textbooks for the 2011-2012 academic year, the electric and magnetism unit category includes electricity and magnetism in the 9th grade, electricity in the 10th grade, magnetism in the 11th grade, and electrical and electronic unit in the 12th grade. In the study of physics textbooks for the 2016-2017 academic year, the material and its properties category includes the material and its properties and heat and temperature in the 9th grade, pressure and buoyant force units in the 10th grade. In the force and motion category, there is force and motion in the 9th and 11th grades, regular circular motion and simple harmonic motion in the 12th grade. The waves category includes waves in the 10th class and wave mechanics in the 12th class. In the modern physics category, introduction to atomic physics and radioactivity, modern physics and technological applications of modern physics

are included in the 12th class. In addition to these, the nature of the physics in the physics textbooks used in the academic year 2011-2012, and the contents of the introduction to science of physics used in the academic year 2015-2016 are the same, that's why they are combined into the introduction to science of physics. In the last stage, the tables were used to put the data into writing.

3. FINDINGS

The first part of this section describes the assessment and evaluation tools used in the physics textbooks used in the 2011-2012 and 2016-2017 academic year, taking into account class levels and units, and the second part shows the assessment and evaluation tools in the same books.

3.1. Distribution of Assessment and Evaluation Tools in the Physics Textbooks by Units

In this section, the assessment and evaluation tools included in the old and new physics textbooks were presented grouped according to the units they are in.

Table 1. Distribution of assessment and evaluation tools in the physics textbooks by units.

	9th Grade						10th Grade						11th Grade				12th Grade					
	The beginning of the unit		Inside the unit		End of the unit		The beginning of the unit		Inside the unit		End of the unit		Inside the unit		End of the unit		Inside the unit		End of the unit		Total	
	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N
Introduction to science of physics	1	2	14	3	25	16											9		10		58	21
Energy		3	20	15	44	18															64	36
Matter and its properties		5	7	16	27	42		3	6	6	19	24	6		25		8		24		122	96
Force and motion		3	21	15	51	26			14		17		19	100	30	144	6	31	27	90	185	409
Electricity magnetism			6		35			3	10	13	21	27	7	68	26	112	5		28		138	223
Waves			11		30			3	8	5	21	27	3		30		12	1	31	40	146	76
Modern physics									7		19		6		26		11	13	25	130	94	143
Stars to quasars													9		11						20	
Atoms to quarks															13		5		16		34	
Optic								3		6		36										45
Total	1	13	79	49	212	102		12	45	30	97	114	50	168	161	256	56	45	161	260	861	1049

O: Physics textbooks used in 2011- 2012 academic year
N: Physics textbooks used in 2015- 2016 academic year

When the total number of questions is taken into consideration as seen in Table 1, the questions in the new textbooks (books in the academic year of 2016-2017) are more than the old textbooks (books in the academic year of 2011-2012). When the units are examined, the most of the questions are in the force and motion unit in both books. Electricity, magnetism and matter and its properties follow this unit. The least of the questions belongs to the unit of stars and quasars and the unit of atoms to quarks in the old textbook, the unit of introduction to science of physics and the unit of energy in the new textbooks.

When the distribution of the questions in the units are examined according to their placements whether they were beginning of the units, in the unit and end of the unit, in the 9th

grades there was only 1 question at the beginning of the unit in the old textbooks. In the new textbooks, there are 13 questions in the 9th grade and 12 questions in the 10th grade at the beginning of the unit. When through the unit evaluation questions are examined, the most of the questions are in the 11th grade of the new textbooks. In the old textbooks, through the unit most of the questions are included in the 9th grade. The number of questions in the other grades is approximately equal. Looking at the end of unit questions, there are more questions in the old textbooks than the new textbooks in the 9th grade only when compared with the textbooks in the two different programs according to the grade level. In other grades, the number of questions in the new textbooks is higher. Especially in the 11th and 12th grades, the number of end of unit questions is higher. The most questions at the end of the unit are at the 12th grade in the new textbooks.

When the unit of force and motion is examined, it appears that only new textbooks of the 10th grade do not include this unit, all other textbooks included it. In this unit, old textbooks do not include questions at the beginning of the unit, while new textbooks make use of questions at the beginning of the unit in grades 9 and 10. When the question distribution of the same unit is examined, the most questions are placed through and at the end of the new textbooks of the 11th grade. The electricity and magnetism unit is another unit that is frequently included in both textbooks and contains many questions.The unit is included in all the classes in the old textbooks while it is not in the 9th and 12th grades in the new textbooks. When the number of questions is examined, it is seen that the most question distribution is through the unit and at the end of the unit in the 11th grade new textbooks. Although matter and its properties unit are included in all grades in the old textbooks, this unit is not available in the 11th and 12th grades in the new textbooks. When the total number of questions belonging to the same unit is examined, the number of questions in the old textbooks is more. When you look at the number of questions by the grades, the most questions about this unit are at the end of the unit of the 9th grade in the new textbooks.

Waves unit is included in all classes in the old textbooks, while it is in 10th and 12th grades in the new textbooks. In the same unit, through unit questions are more in the old textbooks. End of unit questions in this unit have a higher number in the new textbooks. Whereas the modern physics unit was in the 10th, 11th, and 12th grades in the old textbooks, it is only in the 12th grade in the new textbooks. In the new textbooks, 130 questions were found at the end of the unit meanwhile in the old textbooks, there are 94 questions in all units. However, in the old textbooks, the number of questions through the unit and end of the unit is closer to each other. Introduction to science of physics unit is only in the 9th grade in the new textbooks. It is included in the 9th and 12th grades in the old textbooks. In the old textbooks, the only question that is at the beginning of the unit belongs to this unit. However, the number of questions in both units in these two textbooks is considerably less than in other units. The energy unit is only in the 9th grade in both textbooks. The total number of questions in the old textbooks is about close the number of questions in the new textbooks. In the new textbooks, there are about the same number of questions at the end of the unit and through the unit, while in the old textbooks the number of end of unit questions is about close the number of questions through the unit.

Atoms to quarks unit is in 11th and 12th grades in the old textbooks. Whereas in the 11th grade, there are only 13 questions at the end of the unit, in the 12th grade, there are five questions through the unit and 16 questions at the end of the unit. Stars to quasars unit is only in the 11th grade in the old textbooks. Although there are close numbers of questions through the unit and at the end of the unit, this unit is less than the other units in terms of the total number of questions. Optic unit is the only unit that exists in the old textbooks but not in the

new textbooks. This unit is in 10th grade. Although there are a few questions through the unit and at the beginning of the unit, there are more questions at the end of the unit than those.

3.2. Distribution of Assessment and Evaluation Tools in Physics Textbooks

In this section, the assessment and evaluation tools included in the physics textbooks are examined and presented according to their situations in the class and the unit.

As seen in Table 2, when the assessment and evaluation tools in Physics textbooks are examined, mostly open-ended questions are included in the old and new textbooks. When the distribution of this assessment tool is examined, all the questions at the beginning of the unit are in this category. Looking at the questions within the unit, open-ended questions are included in all classes and books, but it appears to be used widely in the 11th grade in the new textbooks. When examining end of unit questions of the same measuring instrument, it was not used at the end of the unit in the 9th and 10th grades in the new textbooks, but it was preferred at the end of the unit in all other books. Multiple-choice questions are the most preferred another assessment tool. This question type is found in all classes only at the end of the unit. The old and new textbooks approximately have the same number of this type, but it is less used in the old textbooks of the 10th class. Gap filling is another assessment tool that is often used in both textbooks. This measuring instrument was used only in the old textbooks at the 9th grade while it was used at the end of the unit in all other classes. True false tests are another assessment and evaluation tool used in the old textbooks of all grades and at the end of the units in the new textbooks of 10th and 11th grades.

Projects have been preferred in all grades and textbooks. Unlike other measuring instruments, however, this measuring instrument is used only through the unit in all books. Research assignments are usually preferred in the new textbooks only through the unit. In the old textbooks, only research studies were included in the 10th grade, whereas this assessment tool was used at all class levels in the new textbooks. Matching questions exist in both textbooks. This measuring tool is used only in the 9th grade in the old textbooks and 9th and 11th grades in the new textbooks. The short answer questions in both textbooks were used through the unit of the 10th grade in the new textbooks which were found in the 9th and 11th grades at the end of the units in the old textbooks. Discussion is another assessment and evaluation technique that exists in both books. This technique has been used in all books through the unit questions. The question type is found in all grades in the old textbooks but only in the 12th grade in the new textbooks. Although the problem solving technique is not used much, both textbooks include it. Despite there is one question through the unit in the old textbooks of the 10th and 12th grades, but there is one question in the 10th grade in the new textbooks at the end of the unit.

Table 2. Distribution of measurement types of assessment and evaluation tools in physics textbooks by classes.

	9th grade						10th grade						11th grade				12th grade					
	At the beginning		Inside		At the end		At the beginning		Inside		At the end		Inside		At the end		Inside		At the end		Total	
	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N
Open-ended	1	13	17	42	43			12	3	21	48		4	152	37	196	1	39	54	60	207	535
Multiple choice					51	53					8	65			45	20			52	60	156	198
Gap-filling			3		40	43					21	49			38	20			30	70	132	182
Meaning analysis table			3		3				2				10				6				24	
Project			30	4					3	4			1	2			2	1			36	11
True false tests					54						13			6	20	20			16	70	103	96
Diagnostic branch tree					5						2				5				3		15	
Table filling			10		4				4								1				19	
Discussion			12						9				5				2	2			28	2
Poster			2						4				7								13	
Pairing					2	6									7						9	6
Short answer					6					4					4						10	4
Problem solving			1																		1	
Information map			1		1																2	
Performance									11				23				42				76	
Discussion																	1				1	
Crossword					2																2	
Concept cartoons					1																1	
Research				3					8	3				4			3				8	13
Concept mapping											5				5				6		16	
Problem solving									1		1						1				2	1
Modelling										1												1

O: Physics textbooks used in the 2011-2012 academic year
N: Physics textbooks used in the 2015-2016 academic year

Although modeling is the only assessment technique that is used in the new textbooks but not used in the old textbooks. This technique is only included in the 10th grade in a question through the unit. Assessment and evaluation techniques included only in the old textbooks are semantic feature analysis, diagnostic branched tree, table filling, posters, information map, performance, debate, puzzles, concept cartoons and concept mapping. Performance tasks are the most preferred of these techniques. This technique has been frequently used through unit questions in grades 10, 11, and 12. Meaning analysis tables were used at the end of the units in all grades, but only in grade 9 it is used through the unit. The diagnostic branched tree was used at the end of the units in all classes. Table filling was found through the unit and at the end of the unit in the 9th grade, while it was never used in the 11th grade. It was preferred at the end of the units in the 10th and 12th grades. Posters were in the 9th, 10th and 11th grade, although they were not in the 12th grade. Concept mapping were at the end of units in grades 10, 11 and 12. The information mapping was used only in the 9th grade through the units and at the end of the units in one question, the debate was used in the 12th grade in one question through the unit, the puzzle was used in the 9th grade only in two questions at the end of the unit and concept cartoon was used only in one question in the 9th grade at the end of the unit.

4. DISCUSSION AND CONCLUSION

When the number of questions is examined, the least number of questions is in the units of stars to quasars and atoms to quarks. The mentioned units do not require much mathematical processing. In the old and new physics textbooks, most of the questions belong to the force and motion unit. In both books, this unit is followed by the units of electricity and magnetism and matter and its properties. It is known that there are many questions that require mathematical operations in these units. The force and motion unit is seen as a unit requiring the most mathematical knowledge by physics teachers (Başkan, Alev &Karal, 2010; Bayrak & Bezen, 2013). It is believed that the high number of questions in these units would allow students to practice more to close the mathematical deficiencies. Yet, Karakuyu (2008) states that students have difficulties to perform mathematical operations in physics classes. Although concept-based teaching is emphasized, It is clear that questions require mathematical processing in physics courses cannot be excluded. This result shows that physics cannot be abstracted from mathematics (Başkan, 2011).

In all textbooks, the number of questions at the beginning of the unit is very few. This number is only one in the beginning of the unit in the old textbooks, and scarcely any in the new textbooks. However, the beginning questions of the unit have an important place in the examination of the students' knowledge and in arousing interest. This is completely ignored in the textbooks. When examining the question distribution in terms of units in the old and new textbooks, the number of end of unit evaluation questions is more than the number of through unit assessment and evaluation. It may be a consequence of the traditional approach being influenced while preparing textbooks. The traditional approach is based on the narrative method and the students are evaluated by end of topic questions. Similarly, it has been pointed out that textbooks are influenced by the traditionalist approach in the study of primary school mathematics books conducted by Arslan and Özpınar (2009). It cannot be expected that the students will go beyond memorization with the courses prepared and the books used according to this approach. Particularly in newly prepared textbooks, the number of end of unit questions is considerably higher than in the old textbooks. It is known that physics teachers do not have enough knowledge about alternative assessment and evaluation techniques and they focus on measuring results rather than process oriented assessment (Akdeniz & Paliç Şadoğlu, 2012). As a result of this situation, it can be thought that the old textbooks did not reach the aim of alternative assessment and evaluation. Teachers may also be focused on evaluating results in

new textbooks because of the feedbacks about the difficulties experienced in implementing the program.

In the new curricula implemented since the 2004-2005 academic year, the traditional assessment and evaluation methods are not completely ignored, and it has been argued that traditional and alternative assessment and evaluation tools should be used together (Yazıcı & Sözbilir, 2014). However, when assessment and evaluation instruments are examined, it is seen that old and new books mostly use open ended, multiple choice, gap filling and true false tests. Besides, the most preferred type of question is open ended questions. Especially in newly written books, the number of open-ended questions is about close that of old textbooks. A similar situation emerged in the study by Çetin and Çakır (2013) of the assessment and evaluation tools in biology textbooks. Ozturk, Yalvaç Hastürk and Demir (2013), in studying the assessment and evaluation approaches used by teachers in science and technology lessons, found that multiple choice and open ended questions were preferred mostly. One reason for this is that open ended questions are one of the most appropriate assessment tools for measuring problem solving and high level skills. Another reason for this may be that the program developed in 2013 ignores the discoveries and experiments and switches to assessment and evaluation centered on the university entrance examination system (Yiğit, 2013). However, the university entrance exam should be based on discovery (Bezen, Bayrak & Aykutlu, 2016). As a result, it is important to remember that students will improve their ability to understand and interpret by moving away from memorization. In contrast, students focus solely on memorizing, and ignore comprehension, practice, and evaluation because of existing books.

One of the goals of alternative assessment and evaluation is to spread the measure to the process instead of a result-oriented approach (Erdoğan, 2007). In the old textbooks, many alternative assessment and evaluation tools such as project assignments, performance task, discussion, concept mapping were included in the unit to reinforce students' learning of concepts. As already mentioned in NTCM (1995), one of the purposes of providing such assessment and evaluation tools in the process is to support learning in addition to revealing the knowledge of students by alternative assessment and evaluation. However, in the physics textbooks prepared in 2016-2017, alternative assessment and evaluation tools which are included in the unit and aimed at process evaluation are given little publicity compared to the old textbooks. One reason for this is that teachers and textbook authors maybe misinterpreted assessment and evaluation as a result of the fact that assessment and evaluation examples are not included in the curriculum developed in 2013 (Göçen & Kabaran, 2013). The 2013 curriculum suggests taking advantage of a variety of assessment methods and indicates them in the program. However, the assessment tools used in the 2016-207 physics textbooks did not go beyond open ended questions, gap filling, true false and multiple choice tests. This may be a sign that new physics textbooks ignore methods that target student centered and alternative assessment.

When we look at the deficiencies in the 2011-2012 physics textbooks, it is seen that some assessment and evaluation tools such as concept cartoons and puzzle are given very little publicity, but some alternative assessment and evaluation tools such as structured grid and word association test have never been used. In addition, concepts such as concept network, concept mapping, concept cartoon, diagnostic branched tree and meaning analysis table have been used always at the end of the unit. Similar findings were also presented by Kavcar (2012). This can be interpreted as the fact that the program does not adequately understand the criterion of the alternative assessment and evaluation, and therefore the necessary value is not given.

Projects, performance tasks and table filling in the alternative assessment and evaluation approaches were frequently used in the 2011-2012 physics textbooks. However, in the 2016-207 physics textbooks, research and questioning based assessment tools were not used

adequately. This situation can cause students to become dependent on the textbook without acquiring knowledge. In addition to this, they can prevent them from going to research and investigation studies other than the textbook. As a result, the textbooks will be confronted only as sources that have adopted the traditional approach of narration and are not used much. If program developers who want their textbooks and curriculums to be implemented and used may attach importance to evaluating the process for the interests and needs of the students, they may be able to close this gap to some extent.

Elimination of the deficiencies in the textbooks is one of the most important studies that increase the quality of education. The incomplete and difficulties in the application are corrected in line with the feedback from the current program and textbooks. However, when the old and new physics textbooks are examined, it can be seen that the deficiencies in the field of alternative assessment and evaluation in the old textbooks have not been solved in the new textbooks, on the contrary, a traditional teacher centered approach has been experienced. As a result, students will come back to memorize again and it will result in that the information will not be used or practiced again.

A successful assessment and evaluation should be at the basis of a successful physics education and this should not be forgotten in the process (Koç & Yayla, 2015). Alternative assessment and evaluation may be a good advantage for physics courses, where success is frequently poor and emphasized by students with negative attitudes. However, if the teachers are not ready for alternative assessment and evaluation, the students are directed to read and memorize because of the content of the questions in the university entrance exam. This situation presents to the students a curriculum of physics lessons that is not parallel to the elementary curriculum exhibiting constructivist and discovery-based instruction and affects their development negatively.

In the light of these results, it should not be forgotten that the university entrance exam has the key role to make students regain the experimenting and discovery which are the essence of physics. Rather than the arrangements to be made in the lessons, it is firstly necessary to regulate the university entrance exams with a student centered structure. Later, it is thought that teachers and students will embrace the student centered approach much more and use it more frequently in their lessons. As a result of this study, it is suggested that researchers analyze the content of the questions asked in the university entrance exam and compare the structure and content of these questions with the data and questions in the current curriculum and textbooks.

Acknowledgement

A part of this study was presented as oral presentation at 13th National Science and Mathematics Education Congress

5. REFERENCES

Akdeniz, A., & Paliç Şadoğlu, G. (2012). Yeni fizik öğretim programına ve uygulanmasına yönelik öğretmen görüşleri [Teachers' opinions about new physics education program and its implementation], *Millî Eğitim Dergisi (National Education Journal)*, 196, 290-307.

Arslan A., Ercan O., & Tekbıyık A. (2012). *Fizik dersi yeni öğretim programına ilişkin öğretmen görüşlerinin çeşitli değişkenler açısından değerlendirilmesi [Assessment of teacher opinions related to physics teaching new curriculum in terms of different variables]*. X. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi (X. National Science and Mathematics Education Congress), Niğde.

Arslan, A., Tekbıyık, A., & Ercan, O. (2012). Fizik ders kitaplarının öğretmen görüşlerine göre değerlendirilmesi [Evaluation of physics textbooks according to teacher views]. *TURJE, 1*(2), 1-13.

Arslan, S., & Özpınar, İ. (2009). İlköğretim 6. sınıf matematik ders kitaplarının öğretim programına uygunluğunun incelenmesi [Evaluation of 6th grade mathematics textbooks along with the teacher opinions]. *Çukurova University Faculty of Education Journal, 36,* 26-38.

Ataman, M., & Kabapınar, Y. (2012). Sosyal bilgiler (4-5. sınıf) ölçme değerlendirme yöntemlerinin kullanılma-kullanılmama nedenleri ve uygulamaların yeterliliği [The purpose and efficiency of the applications of the assessment and evaluation methods in the social studies (4-5th grades) program]. *Amasya Education Journal, 1*(1), 94-114.

Ayvacı, H.Ş., & Devecioğlu, Y. (2013). 10. Sınıf Fizik ders kitabı ve kitaptaki etkinliklerin uygulanabilirliği hakkında öğretmen değerlendirmeleri [Teachers' evaluations on 10th grade physics textbook and applicability of activities in the textbook]. *Amasya Education Journal, 2* (2), 418-450.

Baskan, Z., Alev, N., & Karal, I. S. (2010). Physics and mathematics teachers' ideas about topics that could be related or integrated. *Procedia Social and Behavioural Sciences, 2,* 1558-1562.

Başkan Z., (2011). Doğrusal ve düzlemde hareket ünitelerinin matematiksel modelleme kullanılarak öğretiminin öğretmen adaylarının öğrenmelerine etkileri [The effectiveness of teaching one and two dimensional motion on prospective teachers' learning using mathematical modeling], Doktora tezi (Doctoral Thesis). Karadeniz Teknik Üniversitesi (Karadeniz Technical University), Trabzon.

Bayrak, C., & Bezen, S. (2013). 9. sınıf fizik öğretim programında yer alan konuların öğretiminde karşılaşılan sorunlara ve yeni öğretim programına yönelik öğretmen görüşleri [Teacher opinions on the new teaching syllabus and the issues encountered when teaching the subjects of the 9th grade physics syllabus]. *H. U. Journal of Education,* Özel Sayı (Special Issue) (1), 27-38.

Bezen, S., Bayrak, C., & Aykutlu, I. (2016). Physics teachers' views on teaching the concept of energy. *Eurasian Journal of Educational Research, 64,* 109-124

Çalık, T., (2003). Eğitimde değişimin yönetimi: Kavramsal bir çözümleme [Management of change in education: a conceptual analysis], *Educational Administration: Theory and Practice, 36,* 536- 557.

Çetin, S., & Çakır M., (2013). 2007 Biyoloji öğretim programındaki ölçme ve değerlendirme anlayışının ortaöğretim ders kitaplarına yansımasının değerlendirilmesi [Examining high school biology textbooks in terms of assessment and evaluation approach in the 2007 biology education program], *Trakya University Journal of Education, 3*(2), 104-113.

Çepni, S., (2007). *Araştırma ve proje çalışmalarına giriş [Introduction to research and project studies],* Genişletilmiş 3. Baskı (Expanded 3th Edition), Trabzon: Celepler Matbaacılık.

Çepni, S., Ayvacı, H. Ş., Şenel Çoruhlu, T., & Yamak, S. (2014). Ortaöğretim 9. sınıf fizik ders kitabının güncellenen 2013 öğretim programında yer alan kazanımlara ve kazanımlarda verilen sınırlamalara uygunluğunun araştırılması [An assessment of prospective physics teachers' opinions on the MNE physics textbook for the 10th grade]. *Journal of Turkish Science Education, 11* (2), 137-160.

Eke, C., (2016). Ortaöğretim fizik dersi öğretim programı kazanımlarının webb'in bilgi derinliği seviyelerine göre analizi [Analysis of objectives of high school physics curriculum according to webb's depth of knowledge levels. *Journal of Research in*

Education and Teaching, 5 (3), 35- 40.

Güzel, H., & Adıbelli, S. (2011). 9. sınıf fizik ders kitabının eğitsel, görsel, dil ve anlatım yönünden incelenmesi [Analysis of 9th grade physics coursebook from an educational, visual and language perspective]. *Selçuk University Journal of Institute of Social Sciences, 26,* 201-216.

Göçen, G., & Kabaran, H. (2013). Ortaöğretim 9. sınıf fizik dersi öğretim programlarının tarihsel süreç içerisinde karşılaştırmalı olarak incelenmesi [Comparative study of secondary education 9th grade physics curriculum in the historical process]. *Teaching Science Journal, 1*(2).

Göçer, A. (2008). İlköğretim Türkçe ders kitaplarının ölçme ve değerlendirme açısından incelenmesi [Primary education Turkish course books investigation on measurement and evaluation]. *Atatürk University Journal of Institute of Social Sciences Institude, 11* (1), 197- 210.

Gömleksiz, M. N., Yıldırım, F., & Yetkiner, A. (2011). Hayat bilgisi dersinde alternatif ölçme değerlendirme tekniklerinin kullanımına ilişkin öğretmen görüşleri [Teachers' perceptions of alternative measurement and evaluation techniques used in life sciences classes]. *E-Journal of New World Science Academy, 6*(1), 823-840.

Erdoğan, İ., (2015). *Eğitimde değişim yönetimi [Change management in education]*, 4. Baskı (4th edition), Ankara: Pegem Akademi.

Erdoğan, M. (2007). Yeni geliştirilen dördüncü ve beşinci sınıf fen ve teknoloji ders öğretim programının analizi [Analysis of newly developed fourth and fifth class science and technology curriculum: a qualitative study]. *International Journal of Turkish Education Sciences, 5*(2), 221-254.

İzci, E., Göktaş, Ö., & Şad, S.N. (2014). Öğretmen Adaylarının Alternatif Ölçme Değerlendirmeye İlişkin Görüşleri ve Yeterlilik Algıları [Prospective teachers' views and perceived efficacy regarding alternative measurement and evaluation]. *Journal of Kırşehir Education Faculty, 15*(2), 37-57

Karakuyu, Y. (2008). Fizik öğretmenlerinin fizik eğitiminde karşılaştığı sorunlar: Afyonkarahisar örneği [Problems of physics teachers in physics education: Afyonkarahisar sample]. Mustafa Kemal University Journal of Social Sciences Institute, 5 (10).

Karasar, N., (2007). *Bilimsel araştırma yöntemi [Scientific research method]*, 17. Baskı (17th Edition), Ankara: Nobel Yayıncılık.

Kaya, Ö., (2013). *Yeni fizik dersi öğretim programının ilk yıllardaki uygulamalarına yönelik deneyimlerin incelenmesi [Investigation of Experiences in New Physics Syllabus' First Year Implementation]*, Doktora tezi (Doctoral Thesis). Karadeniz Teknik Üniversitesi (Karadeniz Technical University), Trabzon.

Kavcar, N. (2014). *2013 Ortaöğretim Fizik Programına Uygun Fizik 9 Ders Kitabının İncelenmesi [Assessment of Physics 10 Textbook in Accordance with 2013 Secondary School Physics Curriculum]*, Yayımlanmamış kitap inceleme raporu (Unpublished book assessment report)

Koç, S., & Yayla, A. (2015). Fizik dersi öğretim programının 10. sınıf elektrik ve manyetizma ünitesinin değerlendirilmesi [The physics curriculum evaluation of 10th grade electric and magnetism unit]. *Journal of Research in Education and Teaching,* 4(4).

Kotluk, N., Yayla, A. (2016). Ortaöğretim 9. Sınıf fizik öğretim programının Tyler'ın hedefe dayalı değerlendirme modeline göre değerlendirilmesi [An evaluation of high school 9th

grade physics curriculum according to Tyler's objective based evaluation model], *Abant İzzet Baysal University Journal of Faculty of Education*, 16(4), 1832-1852.

Küçükahmet, L. (2003). *Öğretimde planlama ve değerlendirme [Planning and evaluation in teaching]*, 13. Baskı (13th edition), Ankara: Nobel Yayıncılık.

Küçüközer, H., & Bostan, A. (2007). *İlköğretim 6. sınıf fen ve teknoloji dersi madde ve ısı ünitesinin yapılandırmacı öğrenme kuramının gerekleri ölçüsünde incelenmesi* [*Investigation of elementary 6th grade science and technology course matter and heat unit in accordance with the requirements of constructivist learning theory*]. Ulusal İlköğretim Kongresi *(National Primary Education Congress)*, Ankara.

MEB (2013). *Fizik Öğretim Programı [Physics Teaching Program]*, http://mebk12.meb.gov.tr/meb_iys_dosyalar/31/01/972850/dosyalar/2013_07/05032334_fizik_912.pdf, ET: 14.02.2017.

Nakiboğlu, C. (2009). Deneyimli kimya öğretmenlerinin ortaöğretim kimya ders kitaplarını kullanımlarının incelenmesi [Examination on expert chemistry teachers' secondary school chemistry textbook usage], *Journal of Kırşehir Education Faculty, 10* (1), 91-101.

National Council of Teachers of Mathematics (1995). *Assessment standard for school mathematics*. Reston, VA: NCTM.

Öztürk, N., Yalvaç Hastürk, N.G., & Demir, R. (2013). İlköğretim 4-5. sınıf fen ve teknoloji dersi öğretim programlarındaki ölçme ve değerlendirme yöntemlerine ilişkin öğretmen görüşleri [Teachers' opinions about the assessment and evaluation methods employed in elementary 4-5th grades school science and technology teaching programs]. *Dicle University Journal of Ziya Gökalp Faculty of Education, 20*, 25-36.

Peker, M., & Gülle, M. (2011). Matematik öğretmenlerinin yeni ilköğretim matematik öğretim programında yer alan ölçme araçları hakkındaki bilgi düzeyleri ve bu ölçme araçlarını kullanma sıklıkları [Mathematics teachers' level of knowing about the measurement tools in new elementary school mathematics teaching program and their frequency of use]. *Elementary Education Online, 10*(2): 703-716.

Saban, A. (2009). Çoklu zekâ kuramı ile ilgili Türkçe çalışmaların içerik analizi [Content analysis of turkish studies about the multiple intelligences theory]. *Kuram ve Uygulamada Eğitim Bilimleri Dergisi (Educational sciences: theory & practice), 9* (2), 833-876.

Sağlam Arslan, A., Avcı, N., & İyibil, Ü. (2008). Fizik Öğretmen Adaylarının Alternatif Ölçme Değerlendirme Yöntemlerini Algılama Düzeyleri [Physics Prospective Teachers' Perception Levels Concerning Alternative Evaluations Methods]. *Dicle University Journal of Ziya Gökalp Faculty of Education, 11*, 115–128.

Tabak, R. (2007). *İlköğretim 5. sınıf fen ve teknoloji ders programının öğrenme –öğretme ve ölçme değerlendirme yaklaşımları kapsamında incelenmesi (Muğla ili örneği) [A research about learning teaching and assessment evaluation approaches of the primary school 5. grade science and technology program]*. Yayınlanmamış Yüksek Lisans Tezi (Unpublished Master Thesis), Muğla Üniversitesi (Muğla University), Sosyal Bilimler Enstitüsü (Social Sciences Institute).

Taşdere, A. (2010). *6., 7. ve 8. sınıf fen ve teknoloji ders kitaplarına yansıyan ölçme değerlendirme anlayışının yeni fen ve teknoloji öğretim programı ışığında değerlendirilmesi [The investigation of assessment and evaluation approach that is reflected in 6th, 7th and 8th grades science and technology textbooks prepared in the light of new science and technology teaching program]*. Yayınlanmamış Yüksek Lisans Tezi (Unpublished Master Thesis), Abant İzzet Baysal Üniversitesi (Abant İzzet Baysal University), Sosyal Bilimler Enstitüsü (Social Sciences Institute).

Tay, B. (2013). Sosyal bilgiler öğretmenlerinin alternatif değerlendirme konusundaki görüşleri [The views of social studies teachers about alternative assessment]. *The Journal of Academic Social Science Studies (JASSS), 6*(3), 661-683.

Yazıcı F., & Sözbilir M. (2014). İlköğretim 6-8. sınıf öğretmenlerinin ölçme-değerlendirme yöntemlerine ilişkin kullanım sıklıkları ve yeterlik düzeyleri: Erzurum örneklemi [Elementary 6-8 grades teachers' frequency of use and their level of adequacy in assessment and evaluation methods: Erzurum sampling], *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education, 8*(2), 164-196.

Yıldırım, A. (1999). Nitel araştırma yöntemlerinin temel özellikleri ve eğitim araştırmalarındaki yeri ve önemi [Qualitative research methods]. *Education and Science, 23*, 7-12.

Yiğit, N., Alev, N., Altun, T., Özmen, H., & Akyıldız, S. (2009). Öğretim teknolojileri ve materyal tasarımı, Geliştirilmiş 4. Baskı (Expanded 4th Edition), Trabzon: Akademi Kitabevi.

Yigit, N. (2013). *Ortaöğretim fizik dersi öğretim programı uygulamada ne getirebilir? [What the secondary school physics curriculum can bring in practice?]*. Fen ve Fizik Eğitimi Sempozyumu (Science and Physics Education Symposium), KTU. Trabzon, 26-27 Nisan (26-27April).

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

The Dif Identification in Constructed Response Items Using Partial Credit Model

Heri Retnawati

To cite this article: Retnawati, H. (2018). The Dif Identification in Constructed Response Items Using Partial Credit Model. *International Journal of Assessment Tools in Education*, 5(1), 73-89. DOI: 10.21449/ijate.347956

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 73-89

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

The Dif Identification in Constructed Response Items Using Partial Credit Model

Heri Retnawati[*1]

[1]Mathematics and Science Faculty, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

Abstract: The study was to identify the load, the type and the significance of differential item functioning (DIF) in constructed response item using the partial credit model (PCM). The data in the study were the students' instruments and the students' responses toward the PISA-like test items that had been completed by 386 ninth grade students and 460 tenth grade students who had been about 15 years old in the Province of Yogyakarta Special Region in Indonesia. The analysis toward the item characteristics through the student categorization based on their class was conducted toward the PCM using CONQUEST software. Furthermore, by applying these items characteristics, the researcher draw the category response function (CRF) graphic in order to identify whether the type of DIF content had been in uniform or non-uniform. The significance of DIF was identified by comparing the discrepancy between the difficulty level parameter and the error in the CONQUEST output results. The results of the analysis showed that from 18 items that had been analyzed there were 4 items which had not been identified load DIF, there were 5 items that had been identified containing DIF but not statistically significant and there were 9 items that had been identified containing DIF significantly. The causes of items containing DIF were discussed.

ARTICLE HISTORY

Received: 08 August *2017*
Revised: 23 October 2017
Accepted: 26 October 2017

KEYWORDS

DIF,
polytomous data,
partial credit model,

1. INTRODUCTION

In performing a measurement, there should be utilized valid and reliable instruments. By utilizing instruments that satisfy the both criteria, the measurement results will describe the aspects that should be measured without being influenced by other factors or other loads that should not be measured. An instrument that has been influenced by the other factors other that should be measured certainly contains an error. If the error caused the significance of performance of testees from many groups, it called with bias (Ogbebor & Onuka, 2013).

The bias of a test and a measurement refers to a not good condition, it has unfair meaning, gives to much pressure or becomes too fanatic toward the object under measurement (Osterlind, 1983). The bias within a test has been an unfair and inconsistent condition that has been

*Corresponding Author Phone: +628122774435, E-mail: heri_retnawati@uny.ac.id

ISSN: 2148-7456 online /© 2018 DOI: 10.21449/ijate.347956

contaminated by the factors outside the aspects under the test and by the errors in the test application. This matter shows that the bias within a test and a measurement does not support the characteristics of a valid and consistent test.

Several researchers provide their limitations regarding the item bias, namely Osterlind (1983), Shepard (Adams, 1992), Mazor et al. (1995), Budiono (2004), and Retnawati (2013). A test will be considered biased if two test participants under the same ability from two different groups do not have the same probability to get a correct response. Therefore, the unbiased test items are the ones that have been expected to provide the same probability of providing the correct response among the test participants under the same ability from two different groups (Adams, 1992; Mazor et al. (1995). There are two types of bias namely the external bias and the internal bias.

According to Osterlind (1983), the external bias has been a degree in the test score which shows the correlational relationship of independent variables within a test or an instrument. Furthermore, he states that the problem of the external bias is the social consequence within the test implementation such as the fairness in the test administration and the criteria that might be applied. In relation to this matter, the test administrator has the right to execute the test and to design the criteria that will be related to the fair decisions within the test. Therefore, the aspect that should be given attention within the external bias is the test in overall (the construct validity and predictive validity).

Adams (1992) states that the internal bias which is also known as the item bias refers to the bias within a test that has been related to the psychometric characteristics of a test item and a test in overall. The procedures of detecting the biased items are focused mainly on the investigation whether each test time has similar behaviors or not, namely the similarity in the measurement of psychometric characteristics. According to Osterlind (1983), a test will be considered biased if there is evidence from the interaction between the group members and the test performance in which the different ability or psychological condition among these groups is controlled.

Several psychometric experts have taken the steps to eliminate the lowering connotation in relation to the item bias (Holland & Thayer, 1988; Plake, Patience, & Whitney, 1988). The term that has been used in order to replace the item bias is the differential item performance (DIP) or the differential item functioning (DIF) (Adams, 1992). The new term reflectes the objective of the bias detection method in identifying the items that have different functions for different test participant groups such as the ones that have different facility, different region, different sex and alike.

Based on the results of international studies such as Programme for International Student Assessment (PISA), people can attain information that the literacy scores of Indonesian students has not been satisfying as expected. PISA measures the literacy proficiency that includes the science literacy and the mathematics literacy. These results show that within the conduct of PISA international study the Indonesian students' literacy scores has been far below the international mean (OECD, 2013). Such unsatisfying results might be explored further in relation to the development of the Indonesian students' literacy. Taking a close attention to the test that has been administered by PISA, the respondents of the test are about 15 years old students. These students are both the ones in the ninth grade or in the third grade of junior high school and the ones in tenth grade or the first grade of senior and vocational high school.

The ninth grade students are certainly different than the tenth grade students. The tenth grade students have been provided with the additional materials within the schools, the families and the society for one whole year. These additional materials should be investigated further in order to identify whether they provide additional literacy knowledge or not. In other words,

whether there has been any DIF load or any different probability of providing the correct response toward the test items between the ninth grade students and the tenth grade students or not should be identified. Therefore, this study is to identify the load, the type and the significance of the differential item functioning (DIF) within the partial credit model (PCM) polytomous data. The data that will be manipulated in the study are the students' instrument and the students' response toward the PISA-like test items.

There are several methods that might be applied in order to identify the DIF load within the test items. These methods are classified based on the approach of their underlying theories, namely the classical test theory and the item response theory. In the approach of classical test theory, the methods that have been frequently applied are SIBTES, regression, Mantel-Haenszel (Budiono, 2004), mean covarians (Elosua and Wells, 2013), Lagrange multiplier (Khalid & Glass, 2013) and HGLM (Acara, 2011). Adams (1992) states that the methods that might be applied in order to detect the DIF are factor analysis, item discriminative index by means of point-biserial and partial correlation, item discriminative level test by means of multiple transformations, ANOVA, item response theory or latent trait, chi-square, log-linear model and Mantel-Haenszel statistical theory.

According to Bulut and Suh (2017), there are several methods that might be applied in order to detect the DIF both by means of parametric statistics and of nonparametric statistics. If one would like to apply the parametric statistics methods, then he or she might apply the Chi-Square by Lord, the Likelihood Ratio Test and the Signed and Unsigned Area Methods (Thissen, et al., 1993). On the other hand, if one would like to apply the nonparametric statistics methods then he or she might apply the SIBSTEST or the Mantel-Haenszel methods. The two statements are supported by Retnawati (2003) who performed a DIF analysis using chi-square by Lord and maximum likelihood ratio-test. The methods of both the parametric and the nonparametric statistics might only be applied on a test that measures only one ability (unidimension) and not multiple abilities (multidimension). The existing methods of DIF detection are only found in the unidimension item response theory on the dichotomous score (Camili and Shepard, 1994), the multidimension item response theory on the dichotomous score (Kartowagiran & Retnawati, 2008; Retnawati, 2013) and the likelihood maximum ratio-test (Wang, Yeh, & Yi, 2003).

In the methods of DIF detection by means of item response theory, the DIF is defined as the different probability of providing correct response between two groups that have similar ability. In order to identify the probability difference, the probability of test participants' ability should be identified first. This probability might be identified based on the item parameter, which is adjusted to the scoring type. The test participants' response toward the polytomous scoring-type test items might be analyzed by applying the partial credit model (PCM)-type unidimensional item response theory. At the beginning of the polytomous item response theory development, this model is known more as the expansion of the Rasch model which has been regarded as Partial Credit Model (PCM). The PCM is a polytomous scoring model that has been the expansion of Rasch model in the dichotomous data.

According to Muraki and Bock (1997), the general form of PCM is as follows:

$$P_{jk}(\theta)=\frac{\exp\sum_{v=0}^{k}(\theta-b_{jv})}{\sum_{h=0}^{m}\exp\sum_{v=0}^{k}(\theta-b_{jv})},\ \text{k=0,1,2,...,m} \qquad (1)$$

with:

$P_{jk}(\theta)$ = the probability of θ ability test participants in attaining the *k* score category within the *j* item

θ = test participants' ability

m+1 = the number of *j* item category

b_{jk} = the *k* category difficulty index in the *j* item

$$\sum_{h=0}^{k}(\theta - b_{jh}) \equiv 0 \text{ and } \sum_{h=0}^{h}(\theta - b_{jh}) \equiv \sum_{h=1}^{h}(\theta - b_{jh}) \quad (2)$$

The score of category in the PCM displays the number of the steps that might be taken in order to complete the related test item correctly. The higher score of category resembles the greater ability than that of the lower score of category. In the PCM, if a test item has two categories then the second equation will be the Rasch model equation, like the one that has been proposed by Hambleton and Swaminathan (1985) and that has been supported by Hambleton, Swaminathan and Roger (1991). As a consequence, the PCM might also be implemented toward the polytomous and the dichotomous test items.

In the Rasch model, one of the most famous software for analysis is the QUEST or the CONQUEST by ACER. There are slight differences on the parameter symbols that should be operated. The location parameter between the two software is δ_{ij} instead of *b*. In order to easily understand the related equation and the interpretation of analysis results, the researcher would like to display a mathematical model along the item characteristic curve that is also known as the category response function (CRF).

In order to estimate the parameter along with the *n* test participants (case/person) and the *i* test item with the θ ability and the location parameter of *j* category in the *i* test item that has been equal to δ_{ij} for the 0, 1 and 2 score category, the researcher formulates the following equation (Masters, 2010):

$$P_{ni0} = \frac{1}{\Psi}$$

$$P_{ni1} = \frac{\exp(\theta_n - \delta_{i1})}{\Psi}$$

$$P_{ni2} = \frac{\exp(2\theta_n - \delta_{i1} - \delta_{i2})}{\Psi} \quad (3)$$

Or in general the above equation will be stated as

$$P_{nik} = \frac{\exp(k\theta_n - \delta_{i1} - \delta_{i2} - \cdots - \delta_{ik})}{\Psi} \quad (4)$$

with Ψ as the numerator amount of the overall category.

In the analysis parameter esstimation using a certain software, for example CONQUEST, the δ parameter will be decomposed into the difficulty level parameter and the step parameter. In the 3-category scoring type toward a test item, there will be 2 step parameters and 1 item difficulty parameter. For example, $\delta_{ik} = b_i + \tau_k$ with *b* as the *i* item difficulty parameter and τ as the k step parameter. The probability of each step will be presented as follows.

$$P_{ni0} = \frac{1}{\Psi}$$

$$P_{ni1} = \frac{\exp(\theta_n - b_i + \tau_1\ )}{\Psi}$$

$$P_{ni2} = \frac{\exp(2\theta_n - 2b_i + \tau_1 + \tau_2)}{\Psi}$$

$$\Psi = 1 + \exp(\theta_n - b_i + \tau_1) + \exp(2\theta_n - 2b_i + \tau_1 + \tau_2) \quad (5)$$

The two groups that respond to the test item which has been identified as DIF will be regarded as the focal group and the reference group. The DIF index states the difference of signed area that displays the total probability of providing the correct response in each group. Camilli and Shepard (1994) named this method as Simple Area Indices. Within the test items that have uniform DIF, the DIF index might be identified by:

$$\text{SIGNED-AREA} = \int [P_R(\theta) - P_F(\theta)] d\theta \quad (6)$$

and for the test items that have non-uniform DIF, the DIF index might be identified by:

$$\text{UNSIGNED-AREA} = \sqrt{\int [P_R(\theta) - P_F(\theta)]^2 \, d\theta} \quad (7)$$

By applying the concept of different probability in providing the correct response between the reference group and the focal group, this concept might be applied toward the function of the probability in providing the correct response in the polytomous data. This function is implemented in order to estimate the DIF index that has been developed by Retnawati (2014) by drawing the characteristic curve first. In the test items of polytomous-type test participants' responses that involve two categories, the characteristic curve might be seen in Figure 1.

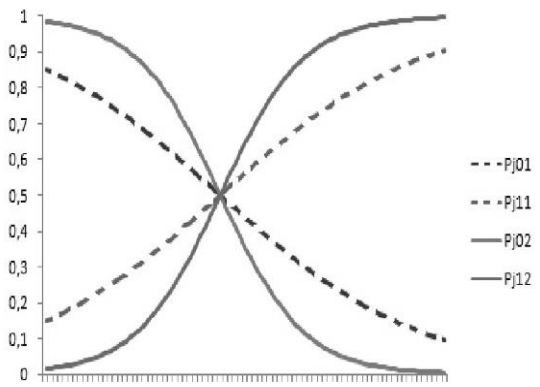


Figure 1.a. The item characteristic curve for the focal (1) a = 0.5 and b = -0.5 and the reference (2) a = 1.2 and b -.05 with 2 categories

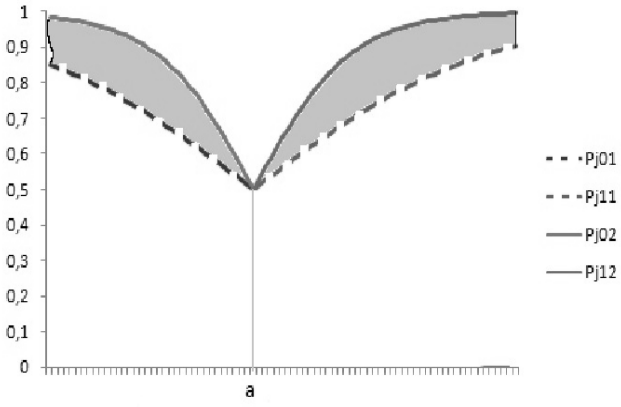


Figure 1.b. The item characteristic curve for calculating the uniform DIF index in PCM with 2 categories

The area between the two characteristic curves is named as the SIGNED AREA, which size might be calculated mathematically by means of integration method. The coverage of this area is the DIF index, which has been drawn in the Figure 1.b. Because in certain points, namely

θ =a, the curve P_{j02} and P_{j12} as well as P_{j01} and P_{j11} are intersecting to each other, the integral equation for the signed area will be:

$$\text{SIGNED-AREA} = \int_{-\sim}^{a}(P_{j02})d\theta + \int_{a}^{c}(P_{j12})d\theta + \int_{c}^{+\sim}(P_{j22})d\theta - \int_{-\sim}^{a}(P_{j01})d\theta - \int_{a}^{b}(P_{j11})d\theta - \int_{b}^{c}(P_{j21})d\theta \quad (8)$$

Similar situation also applies in the 3-category polytomous data that are displayed in Figure 2.a and Figure 2.b. For example, the item parameters of the focal group a = 0.5 are and b_1 = -2.0 and b_2 = 1.0, while the item parameters of the reference group are a = 1.0 and b_1 = 2.0 and b_2 = 1.1. After the item characteristics have been described with the characteristic curve, it is apparent that these items contain the uniform DIF. The coverage of the signed area is formulated through the following equation:

$$\text{SIGNED-AREA} = \int_{-\sim}^{a}(P_{j02})d\theta + \int_{a}^{c}(P_{j12})d\theta + \int_{c}^{+\sim}(P_{j22})d\theta - \int_{-\sim}^{a}(P_{j01})d\theta - \int_{a}^{b}(P_{j11})d\theta - \int_{b}^{c}(P_{j21})d\theta \quad (9)$$

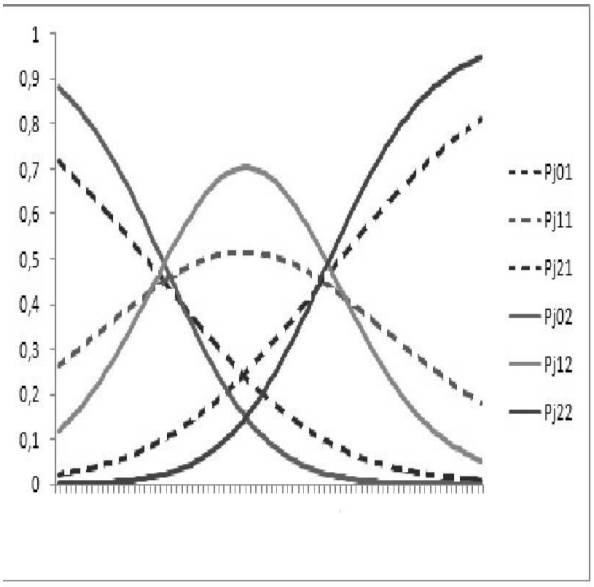


Figure 2.a. The characteristic curve for the focal group (1) and the focal group (2) with 3 categories

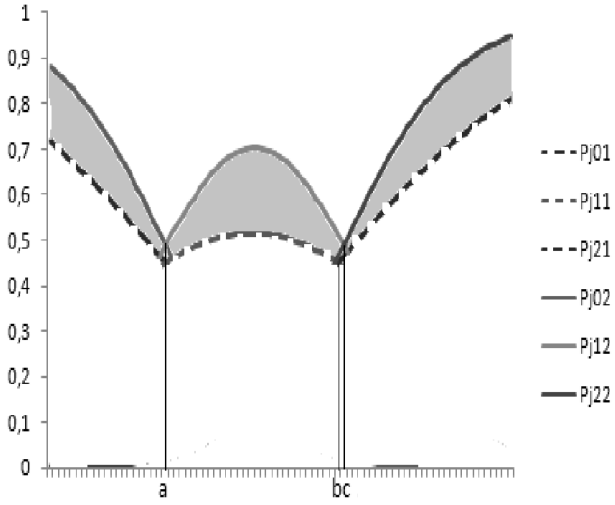


Figure 2.b. The item characteristic curve for calculating the uniform DIF index with 3 categories

In the test items that have non-uniform DIF loads, the DIF index might be identified by paying attention first to the characteristic curve in order to see the integral area. Then, the integral area should be used in calculating the probability coverage. An example of this situation will be provided in Figures 3 and 4.

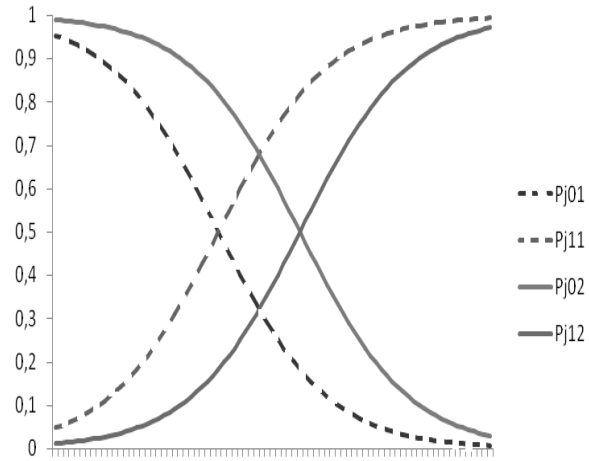


Figure 3.a. The CRF with 2 categories (containing non-uniform DIF loads)

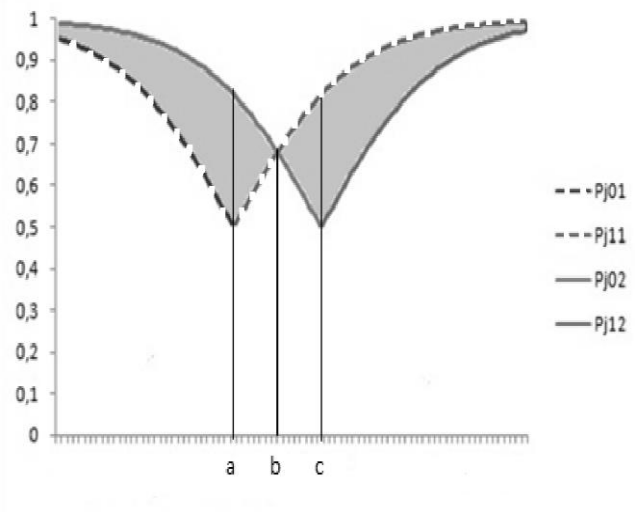


Figure 3.b. Part of the CRF that might be used in calculating the integral of non-uniform DIF loads index in 2 categories

$$\text{UNSIGNED-AREA} = \int_{-\sim}^{a}(P_{j01}-P_{j02})d\theta+\int_{a}^{b}(P_{j01}-P_{j12})d\theta+\int_{b}^{c}(P_{j11}-P_{j02})d\theta+\int_{c}^{+\sim}(P_{j11}-P_{j12})d\theta \quad (10)$$

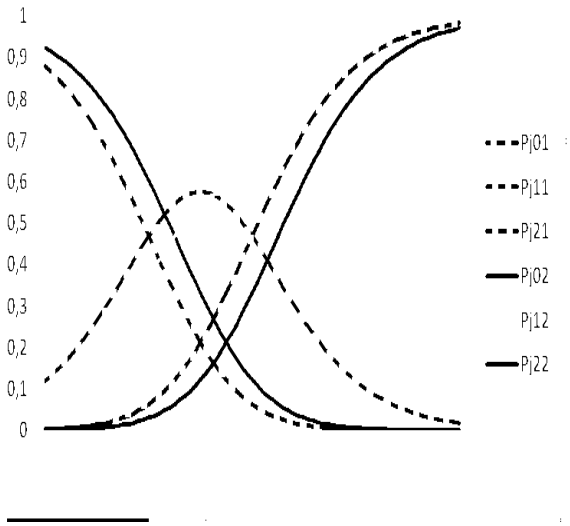


Figure 4.a. The CRF with 3 categories (containing non-uniform DIF loads)

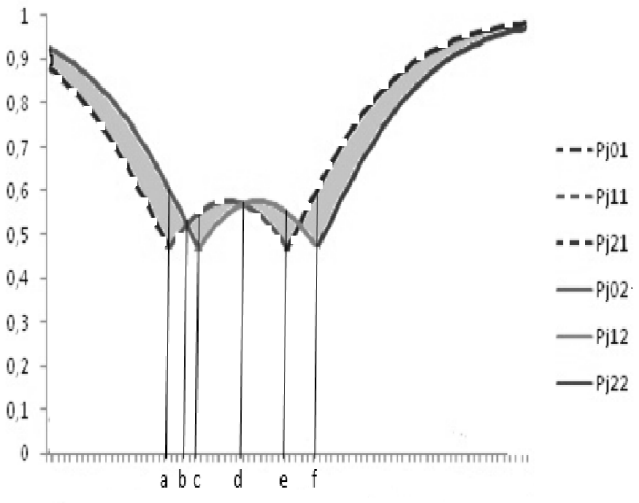


Figure 4.b. Part of the CRF that might be used in calculating the integral of non-uniform DIF loads index in 3 categories

If the function is considered too complicated, the calculation of this integral might be conducted through the Rieman sum calculation assistance by turning the integral area into small area (Varberg & Purchell, 2001) and then calculating these areas by means of numeric approach.

2. METHOD

The study was a descriptive explorative research that identified the DIF loads in the polytomous scoring-type PISA-like test items. The approach in the study was the quantitative one. The study not only identify the load of DIF, but also identify the type and the significance of differential item functioning (DIF) in the partial credit model (PCM) polytomous data.

2.1. Data Collection Method

The data collection of the study utilized test. The test was the PISA-like test instrument that had been developed by Wulandari (Jailani, et al, 2015). The test instrument were developed by adopting the PISA released items from 3 periods (2003, 2007 and 2011); the number of the items was 21 units. The 4 test items had been the constructed responsewith dichotomous scoring (0-1) and 17 test items had been the constructed response with 3 category polytomous scoring (0-1-2).The test contained domain of context (that included the personal context, the societal context, the occupational context and the scientific context) and the domain of process (that included formulate, employ and interpret). The PISA-like test were in *bahasa Indonesia* and utilizing Indonesian contexts.

2.2. The Participants

The test participants of the study are 386 ninth grade students (third grade students of junior high school) and 460 tenth grade students (first grade students of senior high school) whose age were about 15-16 years old. The completion of these items involved the students from 4 regencies and 1 municipality in the Province of Yogyakarta Special Region in Indonesia and these students came from both the state schools and the private schools; the category of these schools are high, moderate and low based on the results of their achievement in the National Examination. The ninth grade students belonged to the focal group, while the tenth grade students belonged to the reference group.

2.3. Data Analysis

The item characteristic analysis utilizing classroom-based student categorization was conducted through the PCM by applying the CONQUEST software (Wu, Adam, and Wilson, 1997). Then, by applying the item characteristics, the researcher draw the category response function (CRF) graphic in order to compare the discrepancy between the item difficulty level and the item error.

The detailed steps in performing the analysis would be given as follows:

1) Estimating the item parameter by means of Rasch model both for the dichotomous data and the polytomous data with the CONQUEST assistance
2) Selecting the fit items by implementing the Rasch model
3) Estimating the item parameters for the ninth grade students' responses and the tenth grade students' responses in the polytomous and the dichotomous data with the CONQUEST assistance
4) Drawing the CRF with the assistance of EXCEL software in order to identify whether the items had been neutral, containing uniform DIF loads or containing non-uniform DIF loads
5) Calculating the DIF index using Rieman sum technique.
6) Determining the DIF significance by comparing the different estimation of item difficulty level parameters and the two group-estimation error with the assistance of CONQUEST program, using criterion an item contains DIF significantly if the discrepancy of the difficulty index is more than twice of its standard error (Adams & Wu, 2010).

7) Interpreting the results of the analysis, including identifying the reasons why the items had been difficult for the students, comparing the substance of the test items and comparing the position of these materials in the curriculum contain within the schools.

3. FINDINGS

The characteristics of the test item instruments were in the form of difficulty level, step parameter and model fitness. The results of the analysis would be displayed in the Table 1.

Table 1. The Overall Item Characteristics and Model Fitness

Item	Category	Difficulty Level	Step 1 Parameter	Step 2 Parameter	MNSQ	Model Fitness
CR113	2	-2.093			1.02	Fit
CR117	2	-1.676			0.91	Fit
CR119	2	-2.275			0.94	Fit
CR127	2	2.092			1.04	Fit
CR203	3	-1.099	0.369	-0.369	1.06	Fit
CR204	3	-0.694	-0.083	0.083	1.02	Fit
CR207	3	1.084	2.702	-2.702	0.55	Fit
CR212	3	-0.074	1.705	-1.705	0.93	Fit
CR214	3	-2.891	1.097	-1.097	0.96	Fit
CR215	3	0.224	0.680	-0.680	1.16	Fit
CR216	3	0.105	0.861	-0.861	0.92	Fit
CR220	3	0.762	-0.675	0.675	0.94	Fit
CR221	3	-0.867	0.799	-0.799	1.19	Fit
CR222	3	-0.948	0.297	-0.297	0.95	Fit
CR223	3	-0.091	0.523	-0.523	1.00	Fit
CR224	3	0.822	1.576	-1.576	0.61	Fit
CR225	3	0.096	-0.901	0.901	0.95	Fit
CR226	3	0.523	-1.205	1.205	1.16	Fit
CR228	3	3.513			0.56	Fit
CR229	3	2.064			0.86	Fit
CR230	3	1.421			0.90	Fit

Based on the results that had been displayed in the Table 1, all items were compatible to the Rasch model. There was a tendency that the items that had 2 scoring categories or more would be easier to compare than those that had polytomous scoring categories. In the last 3 items that are CR228, CR229, CR230 the category parameters did not appear in the analysis results; instead, the difficulty level parameters appeared in the analysis results. The reason was that these items had been responded only by some of the test participants. For the item CR228, only 7.41% of testees got 1 score and none got 2 score. For the item CR230, only 25.53% of testee got 1 score and only 4.26% got 2 score. Then, the three items were excluded from the analysis results.

Furthermore, the researcher estimated the parameters of each item both for the ninth grade students and the tenth grade students. The complete results of the estimation would be

displayed in the Table 2. Based on the results that had been displayed in the Table 2, the researcher found that there had been different parameters between the ninth grade students and the tenth grade students. Although the difference was not prominent, both groups seemed to have different characteristics.

Table 2. The Test Item Parameters that had been Estimated Separately Based on the Data of the Ninth Grade Students and the Tenth Grade Students

Item	Category	Ninth Grade			Tenth Grade		
		Level Difficulty	Step 1 Parameter	Step 2 Parameter	Level Difficulty	Step 1 Parameter	Step 2 Parameter
CR113	2	0.023			-0.023		
CR117	2	0.230			-0.230		
CR119	2	0.847			-0.847		
CR127	2	-0.364			0.364		
CR203	3	0.194	0.606	-0.606	-0.194	0.364	-0.364
CR204	3	0.155	0.186	-0.186	-0.155	-0.087	0.087
CR207	3	0.275	1.017	-1.017	-0.275	2.698	-2.698
CR212	3	-0.124	1.374	-1.374	0.124	1.701	-1.701
CR214	3	0.670	1.310	-1.310	-0.670	1.089	-1.089
CR215	3	0.068	0.817	-0.817	-0.068	0.676	-0.676
CR216	3	0.009	0.798	-0.798	-0.009	0.857	-0.857
CR220	3	-0.040	-0.951	0.951	0.040	-0.679	0.679
CR221	3	0.531	0.644	-0.644	-0.531	0.796	-0.796
CR222	3	0.018	0.040	-0.040	-0.018	0.294	-0.294
CR223	3	0.115	0.269	-0.269	-0.115	0.520	-0.520
CR224	3	0.308	-0.339	0.339	-0.308	1.574	-1.574
CR225	3	-0.084	-0.881	0.881	0.084	-0.905	0.905
CR226	3	0.143	-0.522	0.522	-0.143	-1.208	1.208

Utilizing the item parameters in the Table 2, the researcher might describe the category response function for each item and the researcher might identify whether the DIF loads of an item had been identified or not. Based on the CRF description, the researcher might identify as well whether an item had been beneficial for the ninth grade students or for the tenth grade students. An example of CRF description for the DIF analysis toward several items would be displayed in the Figure 1 until Figure 4.

Also by using the item parameters, the researcher might identify the DIF index by means of integral that had been approached by Rieman sum calculation. The significance of DIF loads might be identified from the comparison between the item parameters discrepancy and the twice of its standard errors that had been calculated by means of CONQUEST. The results of CRF description and the table of DIF identification toward the overall items would be displayed in the Table 3.

Table 3. The Results of DIF Significance Test

Item	Category	Identification of DIF Load Based on the CRF	Type of DIF	DIF Index	Discrepancy on the Difficulty Index	Two-Folded Standard Errors	Significance of DIF Load
CR113	2	Not Loading	-	-	0.046	0.236	-
CR117	2	Loading	Non-Uniform	0.444	0.460	0.246	Significant
CR119	2	Loading	Non-Uniform	1.673	1.694	0.262	Significant
CR127	2	Loading	Non-Uniform	0.703	-0.728	0.964	Not Significant
CR203	3	Loading	Non-Uniform	0.172	0.388	0.172	Significant
CR204	3	Loading	Non-Uniform	0.093	0.310	0.180	Significant
CR207	3	Loading	Non-Uniform	3.081	0.550	0.272	Significant
CR212	3	Loading	Non-Uniform	0.342	-0.248	0.174	Not Significant
CR214	3	Loading	Non-Uniform	0.911	1.340	0.218	Significant
CR215	3	Not Loading	-	-	0.136	0.186	-
CR216	3	Not Loading	-	-	0.018	0.182	-
CR220	3	Loading	Non-Uniform	0.161	-0.080	0.274	Not Significant
CR221	3	Loading	Non-Uniform	0.875	1.062	0.242	Significant
CR222	3	Loading	Non-Uniform	0.250	0.036	0.206	Not Significant
CR223	3	Loading	Non-Uniform	0.554	0.230	0.224	Significant
CR224	3	Loading	Non-Uniform	2.722	0.616	0.460	Significant
CR225	3	Not Loading	-	-	-0.168	0.226	-
CR226	3	Loading	Non-Uniform	0.216	0.286	0.330	Not Significant

From 21 items that had been analyzed, 3 items were excluded from the DIF analysis; as a result, there were 18 items which had been tested. From the overall items and based on the characteristic curve, the researcher attained information that all items had been identified to have the non-uniform DIF loads. From the 18 items, there were 4 items which had not been identified as DIF, there were 5 items that had been identified containing DIF but not statistically significant and there were 9 items that had been identified containing DIF significantly.

Utilizing items paramaters from Table 2, item characteristic curve can be drawn. From its ICC, researcher got information about nature of items, in every category. The categories gave information, wether the step item favored a group of testees. In Figure 5, 6 and 7 explaine the three items with different cases.

The item with the code CR117 had been a test item with a food context that the students commonly read, namely *martabak*. This item had two stimuli namely two types of *martabak*; in the test item, there were two *martabak* with different circular shape and different price but they had the same thickness. These *martabak* would be smeared with the combination of two jam layers and the students, then, were asked to define the amount of the combination.

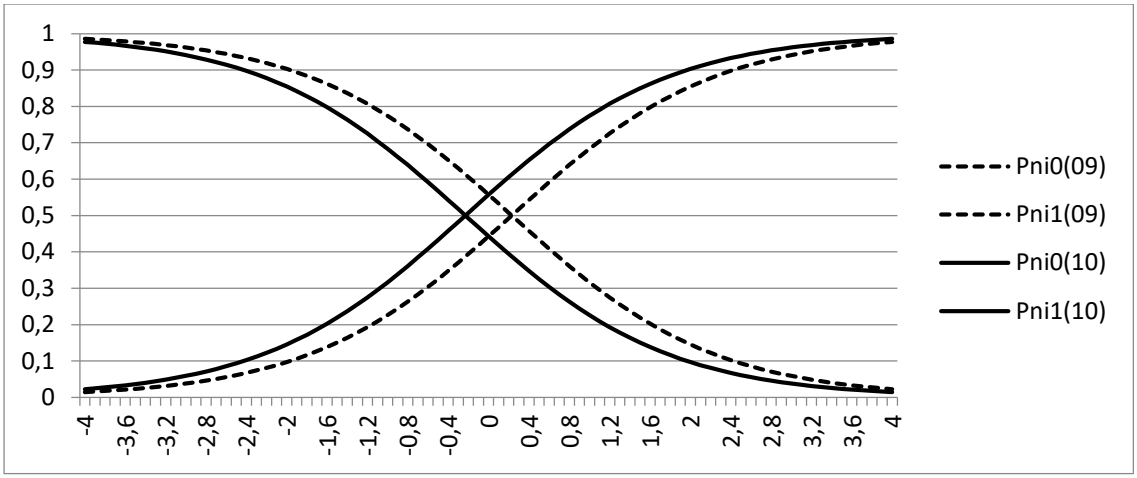


Figure 5. The Graphic of Category Response Function for the Item CR117

Although probability had been studied in the eighth grade, this item demanded specific understanding through the provision of narrative test item. In the item CR117, the tenth grade students had greater chance to score 1 in comparison to the ninth grade students. The reason was that such test items had usually been exercised when the students would attend the national examination; therefore, the tenth grade students, since they used to attend the national examination, would have higher probability in scoring than the ninth grade students.

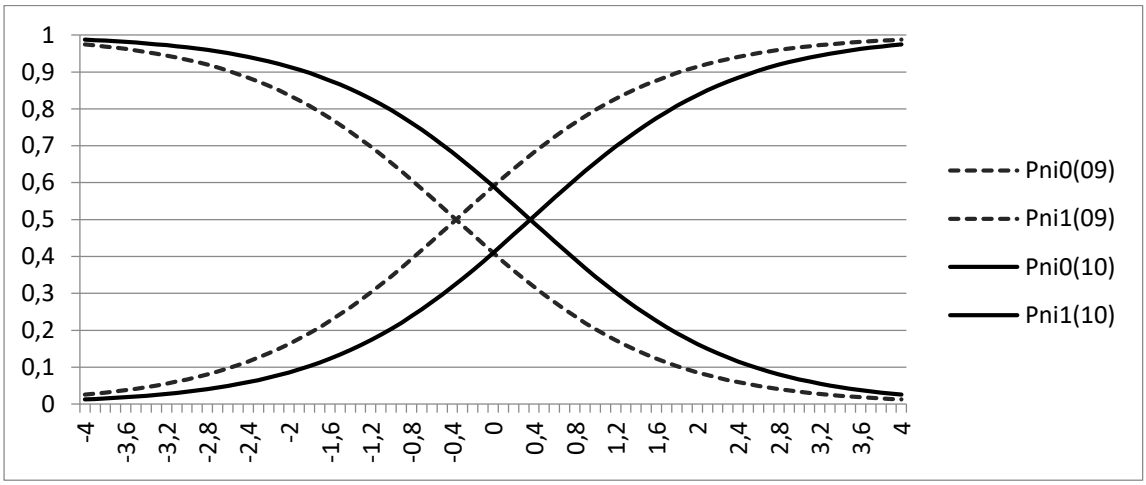


Figure 6. The Graphic of Category Response Function for the item CR127

The item CR127 contained a context where a telecommunication company would like to build a transmitter tower. In this test item, the students were provided with a stimulus of tower construction and of government advice with regards to the construction. Through the concept of distance, the students were asked to provide a reason why the government advice had not been compatible to the regulations of tower construction. The CRF graphic was displayed in the Figure 6. In this item as well, the probability to score 1 among the ninth grade students was higher than that among the tenth grade students. The reason was that the concept of distance had been an easy concept and had been studied much when these students are in the seventh grade. As a result, the ninth grade students had greater probability to memorize this concept than the tenth grade students. It caused the DIF index of the items is equate big, but it is not significantly contain DIF.

The item C212 was beneficial for the tenth grade students both for scoring 1 and scoring 2. This item was related to the materials of probability that had been used in selecting the soccer

players who would take on the penalty shootout and who would have a great probability to be the top scorer. Paying attention to the curriculum that had been applied in the schools, this material was studied by the ninth grade students in their final period. It was the reason why the tenth grade students had higher probability to provide the correct response in order to score 1 or 2. The complete CRF graphic for this item would be displayed in the Figure 7.

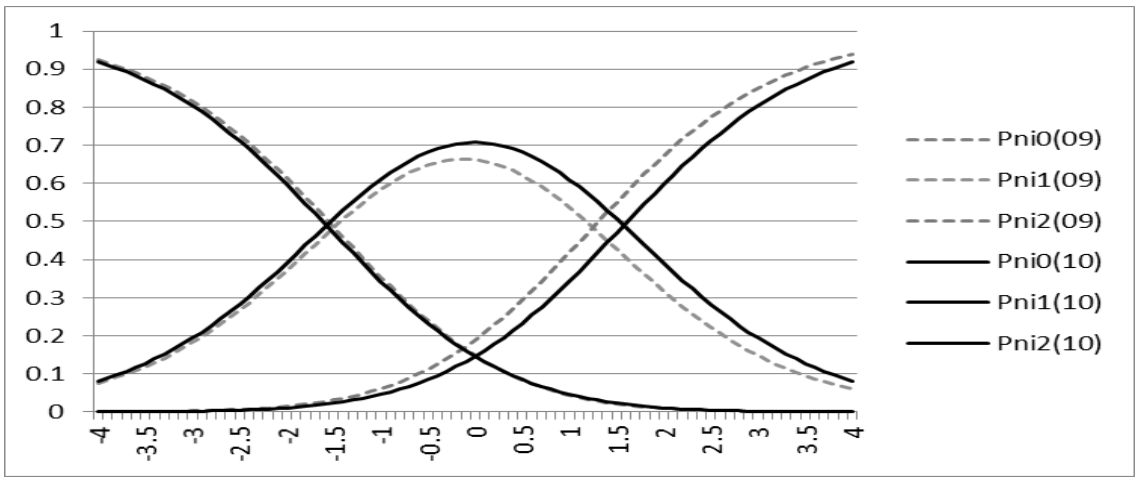


Figure 7. The Graphic of Category Response Function for the Item CR212

A quite different matter was found in the item CR212, which also occurred in the item CR221. The item CR221 had the score 1 category and the tenth grade students had higher probability to score 1 than the ninth grade students. However, in the score 2 category both the ninth grade students and the tenth grade students had the same probability. The reason was that the material in this item had been related to the context of changing the mean values when the test data changed; this material was studied by the ninth grade students in their final period. The CRF graphic for this item would be displayed in the Figure 8.

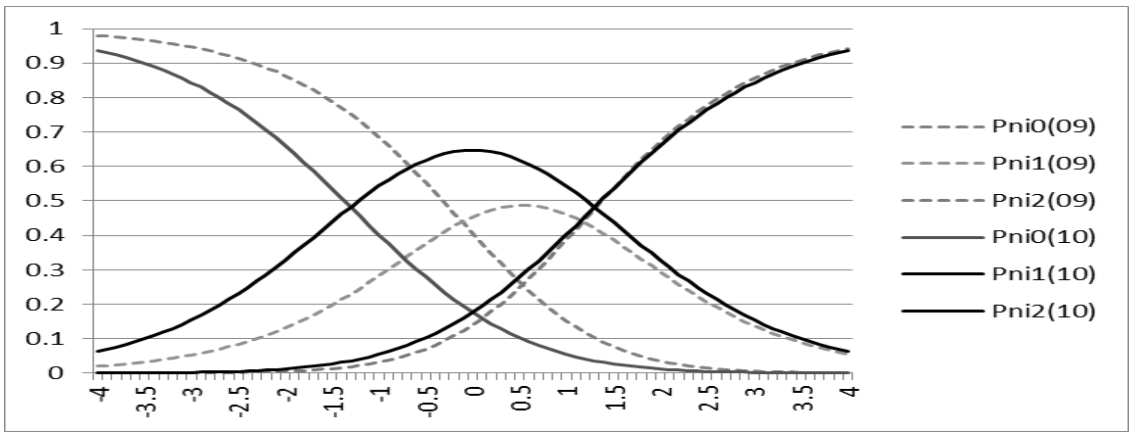


Figure 8. The Graphic of Category Response Function for the Item CR221

The item CR225 had been one of the items that did not have DIF loads. This item had been an item that contained the context of constructing fence in such a way that its circumference would be equal to the length of the wood that the owner had. In order to complete this test item, the students should use their knowledge regarding the concept of determining the circumference of all planes. This material was studied in the elementary school and was

deepened in the seventh grade. Such situation was the reason why the item CR225 did not have any DIF loads.

The item CR225 was also a quite unique item. In the score 1 category curve, the score of maximum probability was lower than the probability score in the intersection of 0 score category and 2 score category. This situation indicated that in this item there had been few students who scored 1 and, as a result, this item might be simplified from 3 answer categories into 2 answer categories. The CRF graphic would be displayed in the Figure 9.

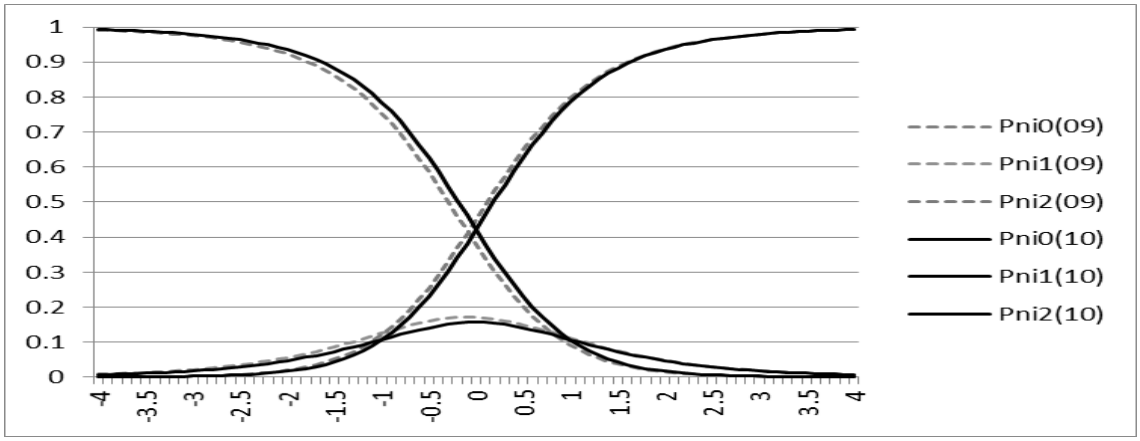


Figure 9. The Graphic of Item Response Category for the Item CR225

The results of DIF significance test in the Table 3 should be given attention as well. By benefitting the estimation resulted-item parameters and the Rieman sum calculation, the researcher attained the DIF index. After the index had been attained, the DIF load significance test was conducted by comparing the discrepancy between the item difficulty level and the parameter estimation errors of the two-group. It turned out that testing the significance through this manner had not been consistent. There were the items which DIF index had been huge but they did not significantly had the DIF loads. On the other hand, there were the items which DIF index had not been huge but they significantly had the DIF loads. In relation to this situation, there should be another study that should pay attention to the comparison in the methods of DIF load identification by using the polytomous data.

Observing each item containing DIF, the most of items contain DIF favoring students aged about 15 years who were in Grade 10, and not favoring students who are about 15 years old but was in grade 9. Based on these results, it can be described the reason why the same age but different classes have the different probability to answer items of PISA-like rightly. The recapitulation of the content and step of items load DIF significantly were showed in Table 4.

Table 4. Recapitulation of content and steps of items load DIF significantly

Item	Content	Step Favore testees from class	
		1	2
CR117	Uncertainty	10	-
CR119	Statistics and Data	10	-
CR203	Geometry	-	10
CR204	Geometry	-	10
CR207	Statistics and Data	10	9
CR214	Uncertainty	10	10
CR221	Statistics and Data	10	-
CR223	Arithmatica	10	-
CR224	Geometry	10	-

4. CONCLUSION AND SUGGESTIONS

The results of the analysis showed that from 18 items that had been analyzed there were 4 items which had not been identified as DIF, there were 5 items that had been identified containing DIF but not statistically significant and there were 9 items that had been identified containing DIF significantly. Many items favored students in grade 9, and another items favored students in grade 10. They were caused by the content of items and depended the posision of the content in the curriculum.

The students aged about 15 years who were in grade 10 had finished studying the subject more than students of the same age, but was in grade 9. It can be seen from the curriculum standards of education in Indonesia (Kementrian Pendidikan Nasional, 2006; 2016). The chapter about statistics and data, and also uncertainty has been learnt by student in the end of 9th, so that those items with this content benefit students in grade 10. Other factor was students of grade 10 has been pass the national exam. Before take this exam, students did a lot of exercises accompanied by deepening material (Sumarno, Sumardiningsih, Muhson, Retnawati, Basuki, 2011). The second thing is what affects the DIF load those polytomous items shaped mathematical literacy is more favor group of participants in grade 10, when compared with a group of students from grade 9. This gives a hint of the development of mathematical literacy skills from grade 9 to grade 10.

The reseach result about DIF load in items of literacy test is in line with many research. The reseach result of Akour, Sabah, and Hammouri (2015) shows that many science items of PISA test contain net and global DIF, and so do in the reading items (da Costa & Araujo, 2012). In mathematics items of PISA, many items in multiple choiche format load DIF favouring male and many items in constructed response load DIF favouring female (Lyons-Thomas, Sandilands, & Ercikan, 2014).

Some future research can be done related to the results of this study. The comparison difficulties of students grade 9 and grade 10 to solve the problems or questions of PISA released items or PISA-like can be done. The development of mathematical literacy skills in grades 9 and 10, or grade level more can be done, either by utilizing the approach of classical test theory and item response theory. Details of students' skills in mathematical literacy, such as domain content, context, and process can be further investigated. The studies result can then be utilized for the improvement of the learning of mathematics.

5. REFERENCES

Acara, T. (2011). Sample size in differential item functioning: An application of hierarchical linear modeling. Kuramve Uygulamada Eğitim Bilimleri (Educational Sciences: Theory & Practice), 11(1), 284-288.

Adams, R.J. (1992). Item Bias. In Keeves, J.P. (Ed), *The IEA technical handbook* (pp. 177-187). The Hague: The International Association for the Evaluation of Educational Achiement (IEA).

Adams, R., & Wu, M. (2010). *Differential Item Functioning*. Retrieved from https://www.acer.org/files/Conquest-Tutorial-6-DifferentialItemFunctioning.pdf

Akour, M., Sabah, S., & Hammouri, H. (2015). Net and global differential item fuctioning in PISA polytomously scored science items: application of the differential step functioning framework. Journal of Psychoeducational Assessment. 33(2), 166-176.

Budiono, B. (2004). Perbandingan metode Mantel-Haenszel, sibtest, regresi logistik, dan perbedaan peluang dalam mendeteksi keberbedaan fungsi butir. *Dissertasion*. Universitas Negeri Yogyakarta, Indonesia.

Bulut, O., & Suh, Y. (2017). Functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. Frontiers in Education, October 2017, 1-14.

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

Da Costa, P.D., & Araujo, L. (2012). Differential item functioning (DIF): What function differently for Immigrant students in PISA 2009 reading items? JRC Scientific and Policy Reports. Luxembourg: European Commission.

Elosua, P., & Wells, C. S. (2013). Detecting dif in polytomous items using MACS, IRT and ordinal logistic regression. Psicológica, 34(2), 327-34

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory*. Boston, MA: Kluwer Inc.

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

Holland, P.W. & Thayer, D.T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In Wainer, Howard; Braun, Henry I. (eds.) Test Validity (p p129-145). Hillsdale, NJ: Lawrence Erlbaum.

Jailani, J., Retnawati, H., Musfiqi, S., Arifin, Z., Riadi, A., Susanto, E., Wulandari, N. F. (2015). Pengembangan perangkat pembelajaran berbasis higher order thinking skills. *Research Report*. LPPM Universitas Negeri Yogyakarta.

Kartowagiran, B. & Retnawati, H. (2008). Pengembangan mengembangkan metode pendeteksian keberfungsian butir pembeda (differential item functioning, DIF) multidimensi. *Laporan Penelitian*. Lembaga Penelitian Universitas Negeri Yogyakarta.

Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2016). Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 21 tahun 2016 tentang Standar Isi. [Ministry of National Education of Republik Indonesia. (2016). *Regulation of Ministry of National Education of Republik Indonesia No 22 Year 2006 about Content Standard in Education*.]

Kementerian Pendidikan Nasional Republik Indonesia. (2006). *Peraturan Menteri Pendidikan Nasional Nomor 22 tahun 2006 tentang Standar Isi*. [Ministry of National Education of Republik Indonesia. (2006). *Regulation of Ministry of National Education of Republik Indonesia No. 22 Year 2006 about Content Standard in Education*.]

Khalid, M.N., & Glass, C.A.W. (2013). A step-wise method for evaluation of differential item functioning. Journal of Applied Quantitative Methods, 8(2), 25-47.

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. Eğitim ve Bilim (Education and Science) 39(172), 20-32.

Masters, G.N. (2010). The partial credit model. In Nering, M.L., & Ostini, R. (Eds). *Handbook of ıtem response theory models*. New York: Routlegde.

Mazor, K. M., Kanjee, A., & Clauser, B. (1995) Using logistic regression and Maentel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32* (2), 131-144.

Muraki, E., & Bock, R.D. (1997). *Parscale 3: IRT based test scoring and item analysis for graded items and rating scales*. Chicago: Scientific Software.

OECD. (2014). *PISA 2012 results: what students know and can do - student performance in mathematics, reading and science*. Paris: OECD Publishing.

Ogbebor, U., & Onuka, A. (2013). Differential item functioning method as an item bias indicator. *Educational Research. 4*(4), 367-373.

Osterlind, S.J. (1983). *Test item bias*. Beverly Hills, CA: Sage Publications Inc.

Plake, B.S., Patience, W.M., & Whitney, D. R. (1988). Differential item performance in mathematics achievement test items: Effect of item arrangement. Educational and Psychological Measurement, 48(4), 885-894.

Retnawati, H. (2003). Keberfungsian butir diferensial pada perangkat tes seleksi masuk smp mata pelalajaran matematika. Jurnal Penelitian dan Evaluasi Pendidikan, 5(6), 45-58.

Retnawati, H. (2013). Pendeteksian keberfungsian butir pembeda dengan indeks volume sederhana berdasarkan teori respons butir multidimensi. Jurnal Penelitian dan Evaluasi Pendidikan, 17(2), 275-286.

Retnawati, H. (2014). Teorı respons butir dan penerapannya. Yogyakarta: Parama.

Salehi, M. & Tayebi, A. (2012). Differential Item Functioning: Implications for Test Validation. Journal of Language Teaching and Research, 3(1), 84-92.

Sumarno, S., Sumardiningsih, S., Muhson, A., Retnawati, H., & Basuki, A. (2013). Faktor yang mempengaruhi menurunnya capaian siswa pada Ujian Nasional 2013. *Laporan Penelitian*. Direktorat PSMP Kementerian Pendidikan Republik Indonesia. [Sumarno, S; Sumardiningsih, S; Muhson, A.; Retnawati, H.; Basuki, A. (2013). Factors affecting students achievement in national examination 2013. *Research report*. Directorate of Secondary School of Ministry Education Office of Republik Indonesia.]

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale NJ: Erlbaum.

Varberg, D. & Purchell, E.J. (2001). *Calculus* (Kalkulus, translated by Susia, I.N.). Bandung: Interaksara.

Wang, W.C., Yeh, Y.L., & Yi, C. (2003). Effect of anchor item methods on differential item functioning detection with the likelihood ratio test. Applied Psychological Measurement. 27(6), 479-498.

Wu, M.L., Adams, R.J., & Wilson, M.R. (1997). *ConQuest: Multi-aspect test software [computer program]*. Camberwell: Australian Council for Educational Research.

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Development of Pamukkale Piano Learning Style Scale

Serkan Demirtaş, Serhat Süral

To cite this article: Demirtaş, S., & Süral, S. (2018). Development of Pamukkale Piano Learning Style Scale. *International Journal of Assessment Tools in Education*, 5(1), 90-104. DOI: 10.21449/ijate.339492

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 90-104

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Development of Pamukkale Piano Learning Style Scale

Serkan Demirtaş, Serhat Süral*

Pamukkale University, Faculty of Education, 20070, Denizli, Turkey

Abstract: In musical instrument training, piano has been taught as a compulsory instrument in all departments of Music Education. It is thought that as a major instrument, piano plays a crucial role in music education. Without question, it is highly vital to raise individuals' awareness of learning styles towards learning piano in effort to practice piano courses more efficiently and effectively. In this respect, the present study is of utmost importance as it will be a pioneer study and make a great deal of contributions to the relevant field. The current study was designed to develop a valid and reliable scale. The population of the study consisted of 170 music teacher candidates majoring in Music Education, including those who already took piano lessons. Although the study successfully accessed to the whole sample, only 133 scales were included to the research, due to inaccurate or incomplete data in subjects' responses. To test the construct validity of the scale, explanatory factor analysis (EFA) and confirmatory factor analysis (CFA) were used. The original scale consists of four sub-dimensions, namely, independent, analytical, dependent and affective learning styles.

ARTICLE HISTORY

Received: 22 September 2017
Revised: 06 November 2017
Accepted: 19 November 2017

KEYWORDS

Learning,
Learning styles,
Individual Differences,
Piano training

1. INTRODUCTION

Individuals living in an age of information are compelled to learn on their own to achieve key elements of learning such as information, skill, attitude and understanding as these learning elements increase and change day by day. In such an age of information in which the information is easily distributed along with the easy access to information, learning and teaching processes should leverage students' individual developments and allow them to adapt innovations. In this context, individual differences should not be ignored and we should strive to find out each student's learning styles and help them to set up a learning infrastructure in their learning process. Today, in modern day education, there is a known fact that what's important is not what a teacher teaches, but how and to what extent a student can learn. An efficient and effective learning will only be achieved as long as such sense of education is adopted. Erden & Akman (2002) highlighted that the one of the critical aspects distinguishing humans from other living creatures is their learning capacity. As biological

*Corresponding Author E-mail: serhatsural@gmail.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.339492**

creatures, humans learn several behaviours in a short time. Firstly, new born humans consciously start to smile to everyone, to learn, to walk and to speak. Then, humans learn to wear, to play with their friends, to read, to write, to play football. Each of them has its own process and each behaviour exhibited in this process is a learned behaviour.

Students are those who achieve learning and all kinds of students' personality traits influence their learning process positively or adversely. Neuropsychological, psychological and physiological aspects of students will shape their future of learning process. Thus, the concept of individual differences becomes prominent. According to Süral (2008); Ryan, (1974); Kulik, (1974); Swanson & Denton (1977), several studies were conducted to investigate how effective individualized teaching was. In previous studies, academic achievements of students who attended courses using direct and critical instruction methods were compared with those of students learning in an individualized teaching system. In this respect, the results revealed that students learning through individualized teaching methods exhibited a high success (Senemoğlu, 2003).

Individualized teaching is a method of teaching in which students do not perform under time pressure; pace of learning is based upon each learner's interest and abilities; individual learning tools, instruments and warning options are delivered to students pertaining to their learning styles; and a continuous feedback is presented to keep students updated about their learning improvements (Tandoğan, 2002).

The concept of individual differences refers to various individual aspects. The very first aspects that come to mind are intelligence, ability and skills, personality traits and learning styles. Individual differences have drawn for many years the attention of the researchers. Educationalists felt the need to explain individual differences. While the concept of individual differences encouraged educationalists to further carry out theoretical studies, individual differences were often neglected in practice. Yet, the fact that each person has a unique character should be considered (Aydoğdu & Kesercioğlu, 2005). As it is known, there is no fixed standard for learning information in the same way. Individuals' learning styles also are different from each other, which should not be ignored and learning environments should be arranged and diversified in this sense. If teaching is performed in such an environment, it will not only contribute to students' academic success but also strengthen their attention span in the learning process. Thus, it is highly vital to identify students' learning styles to achieve these goals. Both teachers and students should be aware of learning styles.

Each person learns in a different way. Each individual is inclined to adopt natural, easy and comfortable learning styles for themselves like the same way they do when they prefer their hairstyles, clothes and food choices. These learning styles allow individualists to effectively access to information with minimum energy and time. Thus, each individual has their own learning styles. As it is an inborn ability, it influences every moment and dimension of human behaviours through their life (Aydoğdu & Kesercioğlu, 2005). Learning style is related to student's individual aspects and preferences. Whereas each individual has unique learning style, they also react to learning. A sense of education in harmony with a student's psychology and environment is the best learning environment for a student (Şimşek, 2007).

Several studies were conducted in the field (Altun, Yurga, Zahal, Gürpınar, 2015; Arslan & Babadoğan, 2005; Aşkar & Akkoyunlu, 1993; Babacan, 2010; Baş & Beyhan, 2013; Bozkurt & Aydoğdu, 2009; Demirtaş, 2017; Duman, 2008; Deniz, 2011; Gencel, 2007; Hasırcı, 2006; Kaleli-Yılmaz, Koparan,; Hancı, 2016; Kaya, Bozaslan, Durdukoca, 2012; Kulaç, Sezik, Aşcı, Gürpınar, 2015; Koçak, 2007; Kolb & Kolb, 2005; Kurtuldu & Aksu, 2015; Okay, 2012; Pehlivan, 2010; Süral, 2008; Sarıtaş & Süral, 2010; Şimşek, 2007; Zahal, 2014;) and many researchers developed learning style models. However, previous studies showed that existing learning styles was based on cognitive success of students or they were

developed to identify individual differences in a general sense. The current study examined learning styles from a different point of view and aimed to find out to what extend learning styles of students talented in art activities were shaped. In this sense, the purpose of the study was to identify learning styles of those individuals talented in playing piano.

As stressed by Say (2001), we can understand piano is important and necessary in music education as a branch of art education. In the phase of musical instrument training, piano has been taught as a compulsory instrument in all departments of Music Education. It is thought that as a major instrument, piano plays a crucial role in music education. Besides, piano is one of the most common instruments used in typical, private and vocational music training. Piano is commonly used because of its high technical capacity, polyphonic feature and broad repertoire (Ömür & Gültek, 2013). As clearly seen, piano will be in the centre of education for an individual who aims to attend fine arts education. Without question, it is highly vital to raise individuals' awareness of learning styles towards learning piano in effort to practice piano courses more efficiently and effectively. In this respect, Pamukkale Piano Learning Styles Model was developed by Demirtaş & Süral to fill the gap in the field.

2. METHOD

The present study was designed to develop a valid and reliable scale.

2.1. Study Group

The population used in this study consisted of 170 music teacher candidates majoring in Music Education, including those who already took piano lessons. Although the study successfully accessed to the whole sample, only 133 scales were included to the research, due to inaccurate or incomplete data in students' responses.

2.2. Data Gathering Instrument

After review of the relevant literature, the scale developed by Karasar (2002) ve Balcı (1995) was selected to use. Accordingly, the following stages were tracked:

1. Pool of Items
2. Expert Opinion
3. Item Analysis
4. Construct Validity of Learning Style Scale
5. Determination of Reliability

The stages mentioned above were outlined as follows:

***Pool of Items*:** In the early stage of scale development process, the following open-ended question was asked of students concerning their thoughts: "What have been your experiences in learning the piano since polyphonic instruments were introduced to you?". The research was administrated to 3rd grade students majoring in Music Education at the Pamukkale University, Faculty of Education, Department of Fine Arts Education.

Item Analysis: The collected compositions were closely reviewed and similar statements were selected. After analysing the statements, scale items were formed and four different learning styles were identified. Afterwards, the scale was called as "Pamukkale Piano Learning Styles Scale (PPLSS)". This study is only applicable to high school and university students due to the sampling group and item content.

Expert Opinion: Experts were consulted to review the item pool. Accordingly, draft scale items were finalized.

Construct Validity of Learning Style Scale: In order to test construct validity of the learning style scale, factor analysis was performed. "Plenty of measurable and observable questions were prepared in an effort to measure psychological aspects of individuals such as

attitude, motive, performance and ability. The question of to what extent scale items measure above-mentioned psychological aspects is related to construct validity" (Büyüköztürk, 2015). Then, the remaining questions were applied to Pamukkale University students in a pilot study. Validity level of the scale were analysed through this pilot study. Therefore, construct validity analysis was carried out via factor analysis technique. After running the factor analysis, four learning styles were determined; 25 out of 55 items were excluded and the original 30 item scale was developed.

Given the scale items measuring learning style, items measuring independent, analytical, dependent and affective learning styles are 1-5-9-13-17-21-25-29, 2-6-10-14-18-22-26, 3-7-11-15-19-23-27-30 and 4-8-12-16-20-24-28, respectively.

Table 1. Reliability Coefficients of the Scale and its sub-dimensions

Factors	Cronbach's Alpha Values
Independent Learning Style	.792
Analytical Learning Style	.792
Dependent Learning Style	.758
Affective Learning Style	.646
Overall	.773

Given the scales are to be used, the level of reliability for preliminary test is expected to be 0.60 as it is 0.80 for fundamental studies. On the other hand, reliability level for practical studies should range between 0.90 and 0.95 (Şencan, 2005). While reliability confidents vary according to types of research in social sciences, reliability confidents for scientific studies are expected to be 0.70 and the level of 0,85 is expected for studies based on ability, interest and skill (Şencan, 2005). All scale items were included and Cronbach's Alpha reliability coefficient of the scale was found to be .773.

2.3. Data Analysis

Initially, draft scale items were transferred into the computer environment according to 133 teacher candidates' responses. The score of each item and the total survey score were calculated. Explanatory factor analysis (EFA) was utilized to test construct validity of the scale and Confirmatory Factor Analysis (CFA) were carried out to evaluate fit indices of the factors obtained. The suitability of the data for factor analysis was determined by running the Kaiser-Meyer-Olkin (KMO) and Bartlett's tests.

3. FINDINGS

Initially, factor analysis was performed using anti-image correlation matrix. The diagonal of anti-image correlation matrix should be greater than .50 (Can, 2014). Items showing a correlation of less than .50 were removed from the survey. The remaining items were subjected to factor analysis. In light of the anti-image correlation matrix results, the diagonal values presented in Table 2 vary between .554 (4th item) and .942 (2nd item).

Table 2. Anti-Image Correlation Matrix

M18	M17	M16	M15	M14	M13	M12	M11	M10	M9	M8	M7	M6	M5	M4	M3	M2	M1	
-,106	,112	-,104	,072	-,035	-,275	,005	,137	-,115	-,001	-,068	-,061	-,124	-,070	,291	-,123	-,459	**,841**[a]	Item 1
,114	-,318	,096	-,112	-,084	,032	,072	,019	-,156	-,215	-,077	-,074	,046	,066	-,084	-,214	**,942**[a]	-,459	Item 2
-,060	,068	-,026	,008	-,153	-,048	-,114	,107	,331	-,111	-,229	,216	-,184	,102	-,375	**,762**[a]	-,214	-,123	Item 3
-,024	-,152	,129	-,171	,071	-,211	-,002	,156	-,404	,034	,349	-,338	,312	-,272	**,554**[a]	-,375	-,084	,291	Item 4
-,181	,014	,141	,050	-,025	,078	,383	-,295	,059	-,205	-,087	-,283	-,019	**,746**[a]	-,272	,102	,066	-,070	Item 5
,020	-,061	-,001	,160	,093	-,041	,206	-,073	-,125	-,192	,020	,037	**,887**[a]	-,019	,312	-,184	,046	-,124	Item 6
,146	,180	-,267	,232	-,044	,066	-,089	-,004	-,027	,027	-,131	**,766**[a]	,037	-,283	-,338	,216	-,074	-,061	Item 7
,080	-,106	,207	-,184	,020	,051	-,129	,040	-,115	-,205	**,881**[a]	-,131	,020	-,087	,349	-,229	-,077	-,068	Item 8
,014	,236	-,087	-,285	-,059	-,300	-,231	,110	-,064	**,897**[a]	-,205	,027	-,192	-,205	,034	-,111	-,215	-,001	Item 9
-,054	,031	-,047	,120	,029	,185	-,067	-,450	**,643**[a]	-,064	-,115	-,027	-,125	,059	-,404	,331	-,156	-,115	Item 10
,060	-,064	,006	-,214	-,013	-,341	-,198	**,590**[a]	-,450	,110	,040	-,004	-,073	-,295	,156	,107	,019	,137	Item 11
,134	-,264	,232	,292	-,285	-,029	**,774**[a]	-,198	-,067	-,231	-,129	-,089	,206	,383	-,002	-,114	,072	,005	Item 12
-,028	-,189	,014	,136	,183	**,869**[a]	-,029	-,341	,185	-,300	,051	,066	-,041	,078	-,211	-,048	,032	-,275	Item 13
-,318	,275	-,272	-,070	**,794**[a]	,183	-,285	-,013	,029	-,059	,020	-,044	,093	-,025	,071	-,153	-,084	-,035	Item 14
,064	-,060	-,422	**,657**[a]	-,070	,136	,292	-,214	,120	-,285	-,184	,232	,160	,050	-,171	,008	-,112	,072	Item 15
-,138	-,230	**,598**[a]	-,422	-,272	,014	,232	,006	-,047	-,087	,207	-,267	-,001	,141	,129	-,026	,096	-,104	Item 16
-,281	**,816**[a]	-,230	-,060	,275	-,189	-,264	-,064	,031	,236	-,106	,180	-,061	,014	-,152	,068	-,318	,112	Item 17
,740[a]	-,281	-,138	,064	-,318	-,028	,134	,060	-,054	,014	,080	,146	,020	-,181	-,024	-,060	,114	-,106	Item 18
-,028	,110	-,239	-,073	,033	-,050	-,130	,062	-,206	,176	-,049	-,060	-,230	,002	-,113	,036	-,005	,139	Item 19
,016	-,078	,123	-,270	,051	-,149	-,166	,284	-,423	-,003	,128	-,063	,030	-,274	,210	-,316	,359	-,088	Item 20
-,344	-,020	,171	-,248	-,103	-,206	-,147	-,031	,056	,082	,175	-,182	-,051	-,028	,087	,003	,028	,000	Item 21
,155	-,114	-,001	,139	-,054	,002	,128	,001	-,132	-,172	-,106	-,009	-,092	,104	,044	-,166	,000	,054	Item 22
-,100	,016	-,106	-,128	,124	-,039	-,345	,159	,120	,091	,118	,006	-,038	-,212	,051	,202	-,149	,012	Item 23
-,026	-,285	-,184	,232	,160	-,285	-,184	,232	,160	-,130	,062	-,206	,176	,359	-,005	,114	-,318	,096	Item 24
,129	-,087	,207	-,267	-,001	-,087	,207	-,267	-,001	-,166	,284	-,423	-,003	-,316	,036	-,060	,068	-,026	Item 25
,141	,236	-,106	,180	-,061	,236	-,106	,180	-,061	-,147	-,031	,056	,082	,210	-,113	-,024	-,152	,129	Item 26
-,001	,014	,080	,146	,020	,014	,080	,146	,020	,128	,001	-,132	-,172	-,274	,002	-,181	,014	,141	Item 27
-,267	,176	-,049	-,060	-,230	,176	-,049	-,060	-,230	-,345	,159	,120	,091	,030	-,230	,020	-,061	-,001	Item 28
,207	-,003	,128	-,063	,030	-,003	,128	-,063	,030	-,130	,062	-,206	,176	-,063	-,060	,146	,180	-,267	Item 29
-,087	,082	,175	-,182	-,051	,082	,175	-,182	-,051	-,166	,284	-,423	-,003	,128	-,063	,180	-,061	,014	Item 30

	M30	M29	M28	M27	M26	M25	M24	M23	M22	M21	M20	M19
M30	,032	,040	-,004	-,073	-,295	,156	,054	,012	,054	,000	-,088	,139
M29	-,084	-,129	-,089	,206	,383	-,002	,000	-,149	,000	,028	,359	-,005
M28	-,112	,051	,066	-,041	,078	-,211	-,166	,202	-,166	,003	-,316	,036
M27	,096	,020	-,044	,093	-,025	,071	,044	,051	,044	,087	,210	-,113
M26	-,318	-,184	,232	,160	,050	-,171	,104	-,212	,104	-,028	-,274	,002
M25	,114	,207	-,267	-,001	,141	,129	-,092	-,038	-,092	-,051	,030	-,230
M24	-,005	-,106	,180	-,061	,014	-,152	-,009	,006	-,009	-,182	-,063	,060
M23	,359	,080	,146	,020	-,181	-,024	-,106	,118	-,106	,175	,128	-,049
M22	,028	-,049	-,060	-,230	,002	-,113	-,172	,091	-,172	,082	-,003	,176
M21	,000	,128	-,063	,030	-,274	,210	-,132	-,120	-,132	,056	-,423	-,206
M20	-,149	-,060	,175	-,051	,028	-,024	-,106	-,118	-,106	-,031	,284	,062
M19	,124	-,106	-,009	-,092	,104	-,073	-,128	-,345	-,128	-,147	-,166	-,130
	-,088	,139	,106	,112	-,104	-,112	,104	,039	,002	-,206	-,149	-,050
	,359	-,005	-,106	-,318	,096	,359	-,054	,124	-,054	-,103	,051	,033
	-,316	,036	-,060	,068	-,026	,008	,139	-,128	,139	-,248	-,270	-,073
	,210	-,113	-,024	-,152	,129	-,171	-,001	-,106	-,001	,171	,123	-,239
	-,274	,002	-,181	,014	,141	,050	-,114	,016	-,114	-,020	-,078	,110
	,030	-,230	,020	-,061	-,001	,160	,155	-,100	,155	-,344	,016	-,028
	-,063	-,060	,146	,180	-,267	,232	-,279	,255	-,279	,154	-,059	**,686[a]**
	,128	-,049	,080	-,106	,207	-,184	,142	-,073	,142	,023	**,727[a]**	-,059
	-,003	,176	,014	,236	-,087	-,285	-,365	,104	-,365	**,853[a]**	,023	,154
	-,423	-,206	-,054	,031	-,047	,120	,417	-,711	**,837[a]**	-,365	,142	-,279
	,284	,062	,060	-,064	,006	-,214	-,051	**,812[a]**	-,711	,104	-,073	,255
	-,009	-,182	-,063	-,060	,146	,180	**,789[a]**	-,115	-,316	,036	-,060	,068
	-,054	-,103	,051	,033	,020	**,814[a]**	-,168	-,156	,210	-,113	-,024	-,152
	-,009	-,092	,104	,044	**,652[a]**	-,073	-,073	,331	-,274	,002	-,181	,014
	-,073	-,295	,156	**,585[a]**	,255	-,100	,016	-,404	,030	-,230	,020	-,061
	-,212	,104	**,714[a]**	,036	-,060	,068	-,026	,059	-,063	-,060	,146	,180
	-,357	**,625[a]**	-,038	-,092	-,051	,030	-,230	-,125	,128	-,049	,080	-,106
	,811[a]	,129	-,087	,207	-,267	-,001	-,087	-,027	-,003	,176	,014	,236

3.1. Construct Validity of the Measurement Tool (Explanatory Factor Analysis)

The suitability of the data for analysis and sampling adequacy was determined by utilizing the Kaiser-Meyer-Olkin (KMO) test. The result of our KMO testis .684 and this value shows that the magnitude of the sample can be characterized as " excellent" for factor analysis and sample adequacy is very high (Kalaycı, 2010; Şencan, 2005; Tavşancıl, 2006;). On the other hand, the results of Bartlett's test indicate that the chi square value (χ^2= 1357.200 ($p< .01$) was significant. In conclusion, the correlation between variables is high. The test results are presented in Table 3.

Table 3. *Kaiser-Meyer-Olkin Test Results*

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.684
Bartlett's Test of Sphericity	Approx. Chi-Square	1357.200
	Degrees of freedom(df)	435
	Sig.	.000

The Varimax rotation technique was performed and items with factor loadings less than .40, items taking place in more than one factor and small items with factor loadings less than 0.10 were extracted from the scale. Yavuz (2005), Bütüner & Gür (2007) proposed that scale

items should not be take place in more than one factor, the criteria for ideal value regarding the difference between the factor loadings should be at least 0.10 and items with factor loadings less than 0.10 should be called as similar items.

Table 4. Factor Loadings of Pamukkale Learning Style Scale

ITEMS	Factors			
	1	2	3	4
Item55	.725			
Item21	.711			
Item29	.642			
Item18	.629			
Item26	.603			
Item53	.573			
Item43	.542			
Item10	.515			
Item7		.750		
Item36		.729		
Item19		.661		
Item39		.641		
Item38		.629		
Item23		.470		
Item15		.420		
Item1		.420		
Item50			.726	
Item52			.716	
Item48			.716	
Item37			.680	
Item2			.637	
Item46			.626	
Item22			.433	
Item32				.742
Item12				.654
Item16				.631
Item28				.583
Item20				.557
Item17				.503
Item47				.422

As the absolute value below was determined as 0.40, values less than .40 was suppressed in items sorted by descending. For this reason, factor loadings given in Table 4 refer to only those factor loadings more than 0.40" (Can, 2014). Factor loadings were determined as 0.40 to make scale items more qualified and distinctive.

Table 5. Eigenvalues of Pamukkale Piano Learning Styles Scale

Factors	*(Initial Eigenvalues)*			*(Extraction Sums of Squared Loadings)*			*Descriptive Statistics*	
	Total	Explained Variance (%)	Cumulative variance (%)	Total	Explained Variance (%)	Cumulative Variance (%)	Mean Factors	Standard Deviation
Independent	4.702	15.672	15.672	4.702	15.672	15.672	38.55	7.263
Analytical	3.536	11.786	27.458	3.536	11.786	27.458	21.22	4.898
Dependent	2.878	9.594	37.052	2.878	9.594	37.052	11.68	3.568
Affective	2.071	6.904	43.956	2.071	6.904	43.956	10.65	2.798

The findings obtained from the factor analysis suggested the presence of four factors with eigenvalues greater than one. Therefore, we can define "Pamukkale Piano Learning Style Scale" as a four-factor Scale. As seen in Table 5, eigenvalues of these four factors and their explained variances were shown. The factors were: "independent learning style" (eight items), "analytical learning style" (seven items), "dependent learning style" (eight items), "affective learning style" (seven items). The eigenvalues of these factors, respectively, are 4.702, 3.536, 2.878 and 2.071 and the results of their explanatory factor analysis demonstrated that these factors, respectively, explained 15.672%, 11.786%, 9.594% and 6.904% of the Pamukkale Learning Style Scale.

It was determined from the explanatory factor analysis (EFA) that these extracted four factors explained 43.956% of the total variance. Şencan (2005) and Can (2014) argued that this variance rate is acceptable. Pearson correlation coefficients were calculated to investigate the relation of the four factors to each other and to the total scale score and the results are shown in Table 6. Based on the findings presented in Table 2, we see that the relation of the four factors to each other and to the total scale score was found significant. Depending on the correlation coefficients of the scale, its reliability is characterized as follows: if it ranges between 0.70 - 1.00, the reliability of the scale is highly reliable; if it ranges between 0.69 - 0.30, the reliability of the scale is moderately reliable; if it ranges between 0.29-0.00, the reliability is low (Büyüköztürk, 2006).

Table 6. Correlation of the four factors with each other and total scale

Factors	Factor 1	Factor 2	Factor 3	Factor 4	Total
Independent L.S. (F1)	*				
Analytical L.S. (F2)	.711	*			
Dependent L.S. (F3)	.687	.654	*		
Affective L.S. (F4)	.598	.705	.688	*	
Total	.857	.811	.768	.741	*

* All correlations have p< 0.01

According to the correlation analysis of four factors with each other and total scale, the correlation coefficients between total score and each factors were determined as follows: "independent learning style" (factor 1) sub-dimension is r= .857; "analytical learning style" (factor 2) subdimension is r= .811; "dependent learning style" (factor 3) sub-dimension is r=.768 and affective learning style (factor 4) sub-dimension is r= .741. Consequently, the fact that the relation between the four factors in the scale and total scale is highly significant

supports the construct validity of the Pamukkale Learning Styles Scale. The results of the KMP and Bartlett's tests were supported as well.

3.2. Language Validity of Pamukkale Piano Learning Style Scale

Pamukkale Piano Learning Style Scale is 5-likert scale of 30 items composed of four sub-dimensions. In this context, independent and affective learning styles consist of eight items and dependent and analytical learning styles consist of seven items. The scale was adapted to English language by three-people team. Afterwards, four out of eight-people group majored in English Literature and Language was asked to translate English items to Turkish and the rest of the group were asked to translate Turkish items to English. As a result of the findings obtained, the scale was finalized in English. Then, English version of the scale was administrated to 60 students majoring in English Teaching. After 10 days passed, the Turkish version of the scale was carried out and the relationship between two versions was compared. In light of the data obtained, significance level was determined using Pearson's Product Moment Correlation Coefficient test. In this context, the significance level was calculated as .714.

Table 7. Explanatory Factor Analysis

Fit Indices	Fit Range	Research Model Four-Factors Model
Total Fit Index		
χ^2/sd	$0 \leq \chi^2/sd \leq 3$	522.17 / 217= 2.40
Comparative Fit Index		
NFI	.90 ≥ - ≥ .94	.92
NNFI	.90 ≥ - ≥ .94	.91
IFI	.90 ≥ - ≥ .94	.91
CFI	≥ ,95	.95
RMSEA	0.05 ≤ - ≤ 0.08	0.071
Absolute Fit Indices		
GFI	≥ .90	.90
AGFI	≥ .85	.85
Residual Based Indexes of Compliance		
SRMR	.06 ≤ - ≤ .08	.069
RMR		.074

As seen in Table 7 to test the reliability of the four sub-dimensions identified through explanatory factor analysis, a confirmatory analysis was performed. Results from confirmatory factor analysis indicated that chi-square was (χ^2=522.17), degree of freedom (df=217, p=0.00) was χ^2/df=2.40, SRMR= .069, RMR=.074, AGFI= .85, GFI=.90, RMSEA= 0,071, CFI=.95, NNFI=.91, NFI=.92, IFI=.91. CFA revealed that χ2 /df ratio is lower than 3. Other goodness for fit indices computed by CFA were: IFI= .90 ≥ - ≥ .94; NFI = .90 ≥ - ≥ .94; NNFI =.90 ≥ - ≥ .94; CFI= ≥ .95; RMSEA= 0.05 ≤ - ≤ 0.08 and GFI= ≥ .90 AGFI =≥ .85 and lastly SRMR and RMR = .06 ≤ - ≤ .08. Consequently, the values mentioned above indicate acceptable fit.

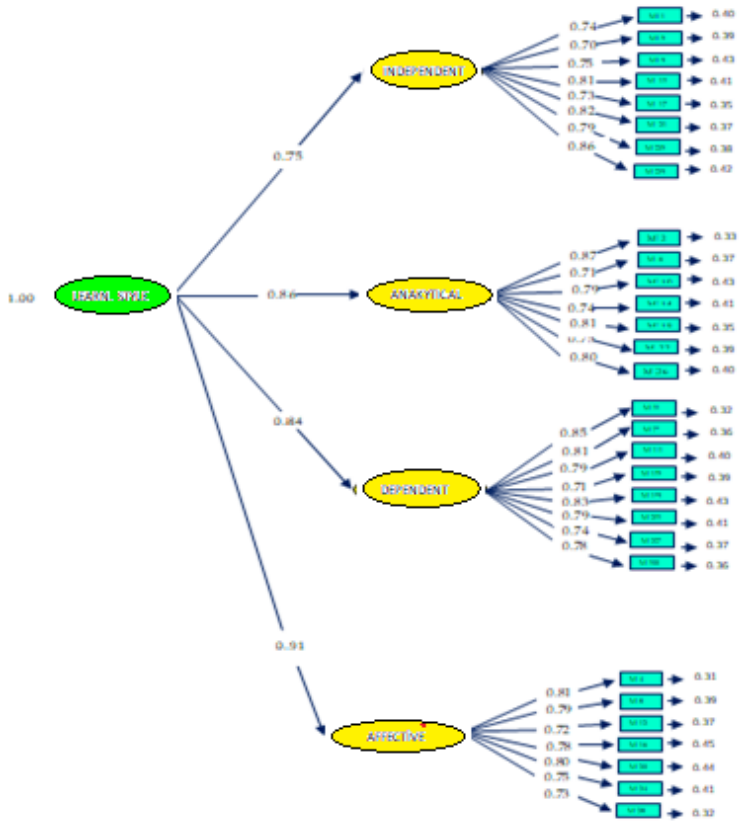


Figure 1. PPÖSÖ Four-Factor Path Diagram

From this data, it can be said that four dimensional constructions about Pamukkale piano learning style scale is appropriate. Substance factor coefficients calculated by confirmatory factor analysis are presented in Figure 1. According to this, item factor direct correlation coefficients ranged from .70 to .87. The error variances of the items ranged from .31 to .45. The observed item was found to be significant in scale relations.

4. RESULTS

As a result of the findings obtained, a learning style model was developed to find out learning style of students playing piano. According to the model, it was understood that students used four different learning styles while learning the piano. These four learning styles were named as “independent”, “analytical”, “dependent” and “affective”.

It was observed that students who prefer independent learning style are individual learners. They don’t need any external factor, a teacher or a friend. Such students can categorize pieces of music they practice, analyse and interpret them from their own point of views. They prefer to learn on their own and exhibit high self-confidence. However, since an individual learner will not benefit from a teacher experience or knowledge, independent learning style can have some drawbacks in terms of students’ vocational experience and performance.

Students who prefer analytical learning style adopt a conceptual view. They don’t work pieces of music as a whole, divide them into sections. Students try different methods and adopt solution-oriented approach in an effort to reach a solution. They prefer individual learning as well. Such students like to work in safe learning environments and they like to divide their works into smaller parts by analysing challenges they encounter. They are good at reading musical scores. They can decipher musical notation quickly. Such students learn in a

planned way and thereby learn pieces more systematically and faster. This can be seen as an advantage in students' learning process. Yet, when students work musical pieces as a whole, they can barely finish playing in time and they are delayed due to passage works, which is seen as a disadvantage in terms of analytical learning style.

Students in a dependent learning group wait for an external warning. Guidance of someone else comforts students and makes students work better when they organize their studies. As such students always are looking for other resources; they cannot read the musical notation very well. When they start to decipher a new notation, they first need to hear it from someone else. They always consult their works to be checked by someone else. In the stage of working on a musical piece, they try to reach audiovisual resources and they play them by imitating. A student using a dependent learning style has a more artistic and musical character as they access to various resources. On the other hand, they have lower self-confidence as they depend on an external factor and they cannot read the notation very well. They complete a musical piece of work in a longer period.

A student adopting affective learning style looks for a familiar tune in a musical piece. Such students can better work if they like pieces of music they play. If they don't like musical piece, they cannot perform effectively. They mostly prefer to play their pieces over and over in a wholly way. They always expect to take positive feedbacks during piano courses and if they take a negative feedback, they alienate themselves from the course. Such students who play their preferred melodies and pieces can easily learn as they have high levels of motivation. They can be successful when they find suitable conditions for themselves. On the other hand, as they always demand to play their favourite pieces, we cannot expect an efficient and qualified training. Students adopting affective learning style cannot accept their teachers' criticism.

5. REFERENCES

Altun, F., Yurga, C., Zahal, O., & Gürpınar, E. (2015). Music teacher candidates' learning styles and the relationship between field achievement points. *E-International Journal of Educational Research*, 6(3), 46-70.

Arslan, B. ve Babadoğan, C. (2005). Relationships between learning style preferences and gender, age and success level at 7th and 8th Grade. *Eurasian Journal of Educational Research,* 21(11), 35-48.

Aşkar, P., & Akkoyunlu, B. (1993). Kolb learning style inventory. *Education and Science,* 87, 37-47.

Aydoğdu, M., ve Kesercioğlu, T. (2005). *İlköğretimde fen ve teknoloji öğretimi*. İstanbul: Anı Yayıncılık.

Babacan, E. (2010). *The Application of Perceptual Learning Styles in Beginning Piano Education.* Unpublished Doctorate Thesis. *Konya:* Selçuk University The Institute of Education Sciences.

Balcı, A. (1995). *Sosyal bilimlerde araştırma yöntem, teknik ve ilkeler.* Ankara: Pegem Yayıncılık.

Baş, G. & Beyhan, Ö. (2013). Effects of learning styles based instruction on academic achievement, retention level and attitudes towards English course. *Ankara University, Journal of Faculty of Educational Sciences,* 46(2), 133-158.

Bütüner, Ö.S., & Gür, H. (2007). Development of an attitude scale for the V diagram. *Journal of National Education*, 176, Spring, 72–85.

Büyüköztürk, Ş. (2006). Sosyal Bilimler İçin Veri Analizi El Kitabı. Ankara: Pegem Akademi.

Büyüköztürk, Ş. (2015). *Sosyal bilimler için veri analizi el kitabı.* Ankara: Pegem Akademi.

Bozkurt, O., & Aydoğdu, M. (2009). A comparative analysis of the effect of Dunn and Dunn learning styles model and traditional teaching method on 6th grade students' achievement levels and attitudes in science education lesson. *İlköğretim Online*, 8(3), 741-754.

Can, A. (2014). *SPSS ile bilimsel araştırma sürecinde nicel veri analizi.* Ankara: Pegem Akademi.

Demirtaş, S. (2017). *The relationship between learning the styles of pre-service music teachers and academic achievement.* Unpublished Doctorate Thesis. Uludağ University, The Institute of Social Sciences.

Deniz, J. (2011). *Learning Styles of Music teacher candidates.* 2. International Conference of New Trends in Education and Their Implications, 950-956

Duman, B. (2008). Comparison of learning strategies and learning styles that students use with their educational philosophy. *Journal of the Cukurova University Institute of Social Sciences*, 17(1), 203-224.

Erden, M., Akman, Y. (2002). *Gelişim ve Öğrenme*. Arkadaş Yayınevi. Ankara.

Gencel, İ.E. (2007). Kolb's learning styles inventory based on experiential learning theory-III adaptation to Turkish. *Dokuz Eylül University The Journal of Social Sciences Institute*, 9(2), 120-139.

Hasırcı, Ö.K. (2006). Learning styles of prospective primary school teachers: Çukurova University Case. *Journal of Theory and Practice in Education*, 2(1), 15-25.

Kalaycı, E. (2010). *Investigation of the relationship between cyber ailment behaviors and self-regulatory strategies of university students*. Unpublished Doctorate Thesis. Hacettepe University Institute of Sciences, Ankara.

Kaleli-Yılmaz, G.; Koparan, T., & Hancı, A. (2016). Determination of the relationship between 8th grade students learning styles and TIMSS mathematics achievement. *Bayburt University Journal of Faculty of Education*, 11(1), 35-58.

Karasar, N. (2002). *Bilimsel Araştırma Yöntemleri.* Ankara: Nobel Yayın Dağıtım.

Kaya, A.; Bozaslan, H., & Durdukoca, Ş. F. (2012). Examination of the relationship between teacher candidates' learning styles and their study habits. *Electronic Journal of Social Sciences*, 11(41), 131-146.

Koçak, T. (2007). *Primary education 6.7.8. An examination of the relationship between classroom learning styles and academic achievement.* Unpublished Master's Thesis. Gaziantep Universtiy Institute of Social Sciences.

Kolb, A. Y.; Kolb, D. A. (2005). Learning styles and learning spaces: enhancing experiential learning in higher education. *Academy of Management Learning & Education*, 4(2), 193-212.

Kulaç, E.; Seziç, M., Aşçı, H., & Gürpınar, E. (2015). Learning styles academic achievement and gender in a medical school setting. *Journal of Clinical and Analytical Medicine*, 6, 5, 608-611.

Kulik, J. A., & Kulik, C. L. C. (1974). Student ratings of instruction. *Teaching of Psychology*, *1*(2), 51-57.

Kurtuldu, M. K., & Aksu, C. (2015). Evaluation on learning styles of candidate music teachers according various variables. *The Journal of Art Education*, 3(2), 1-23.

Okay, H. H. (2012). The relations between academic achievement in field lessons and learning styles of music teacher candidates. *Procedia - Social and Behavioural*

Sciences, 51, 193-197.

Ömür, Ö., & Gültek, B. (2013). Mental processes affecting the piano performance. *International Journal Human of Human Sciences*, 10(1), 417-433.

Pehlivan, K. B. (2010). A study on prospective teachers' learning styles and their attitudes toward teaching profession. *Elementary Education Online*, 9(2), 749-763.

Ryan, W. J. (1974). Perceptual and acoustic correlates of aging in the speech of males. *Journal of communication disorders*, *7*(2), 181-192.

Say, A. (2001). *Müziğin Kitabı*. Ankara: Müzik Ansiklopedisi Yayınları.

Sarıtaş, E., & Süral, S. (2010). Grasha - reichmann learning and teaching style of the scale study turkish adaptation. *E-Journal of New World Sciences Academy,* 5*(*4), 2162-2177.

Senemoğlu, N. (2003). *Gelişim Öğrenme ve Öğretim Kuramdan Uygulamaya***.** Gazi Kitapevi. Ankara.

Süral, S. (2008). *The relationship between learning styles and academic achievements of primary candidate teachers in science and technology teaching lesson.* Unpublished Master's Thesis. Pamukkale University, Institute of Social Sciences.

Swanson, D. H., & Denton, J. J. (1977). Learning for mastery versus personalized system of instruction: A comparison of remediation strategies with secondary school chemistry students. *Journal of Research in Science Teaching*, *14*(6), 515-524.

Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenirlik ve geçerlilik*. Ankara: Seçkin Yayıncılık.

Şimşek, Ö. (2007). *Development of Marmara learning style scale and examination of learning styles of 9-11-year-old children.* Unpublished Doctorate Thesis. İstanbul University, Institute of Education Sciences.

Tandoğan, M. (2002). *Öğretmen ve Teknoloji***.** Anadolu Üniversitesi Açıköğretim Fakültesi Yayınları. Eskişehir.

Tavşancil , E. (2006). *Tutumların Ölçülmesi ve SPSS İle Veri Analizi*. (3. Baskı). Ankara: Nobel Yayınları.

Yavuz, S. (2005). Developing a technology attitude scale for pre-service chemistry teachers, *The Turkish Online Journal of Educational Technology*, 4, 1–9.

Zahal, O. (2014). *Relation between learning styles and cognitive flexibility with examination achievement who enter the special ability exam.* Unpublished Doctorate Thesis. Inonu University, Institute of Education Sciences.

APENDIX 1. Pamukkale Piano Learning Styles Scale

	ITEMS	Strongly Disagree	Disagree	Moderately Agree	Agree	Strongly Agree
1	When I learn a new piece of music, I try to find out the period of the piece and its background and then study accordingly.					
2	Playing a piece in 2/2 measure allows faster progress for me.					
3	It is easier for me to play a piece after I hear it from a friend of mine for the first time.					
4	I love to practice my favourite melodies on the piano.					
5	When I learn a new piece of music, I always examine composers' characteristics.					
6	I practice passage by breaking up a musical paragraph into smaller group of notes.					
7	I try to play musical pieces by ear rather than reading notes.					
8	I can be a quick learner if I have a chance to practice my favourite piece of work.					
9	I prefer to use metronome for piano practice.					
10	I practice piano by splitting musical pieces into staves.					
11	I get motivated to play a piece after I hear it from a friend of mine.					
12	If lecturers make us to love piano lessons, we study harder and learn better.					
13	When practicing piano, I pay attention to work a piece phrase by phrase.					
14	I go through a musical pieces phrase by phrase and then combine them.					
15	As I don't read sheet music very well, I prefer to memorize a piano piece.					
16	I get motivated if I like the melody of a piece.					
17	I certainly pay attention to nuances of a musical work.					
18	When I learn a new piece, I divide it into measures.					
19	I feel confident enough to practice piano only after I hear a piece from someone else.					
20	I always learn faster if I like piano lessons.					
21	I do finger exercising before playing piano.					
22	I always try to divide a piece into 4/4 measure.					
23	To check myself before class, I perform in front of a friend of mine and ask my friend's opinion about my performance.					
24	I firstly analyse a piece and then consider its level of difficulty.					
25	When a new piece of music is assigned to me, I always analyse its harmonic structure.					
26	When I learn a new piece of music, I work on my right and left hands separately.					
27	I try to play pieces by imitating other's works.					
28	When practicing, I mostly repeat a piece over and over again.					
29	It is important for me to decipher notation by using finger numbers.					
30	I always try to memorize notation.					

Turkish version of the scale

APENDIX 2. Pamukkale Piano Öğrenme Stili Ölçeği

	MADDELER	Hiç Katılmıyorum	Katılmıyorum	Az Katılıyorum	Katılıyorum	Tamamen Katılıyorum
1	Yeni bir parça çalışırken o parçanın hangi döneme ait olduğuna bakıp o dönemin özelliklerini öğrenerek çalışırım.					
2	Parçalarımı ikişer ölçü biçiminde çalışmak beni daha hızlı ilerletir.					
3	Parçalarımı başka bir arkadaşımdan dinlemek daha kolay çalışmamı sağlar.					
4	Hoşuma giden melodileri çalışmayı isterim.					
5	Çalışacağım eserin bestecisinin özellikleri hakkında inceleme yapıp araştırırım.					
6	Çalıştığım parçayı küçük birimlere bölerek pasaj çalışması yaparım.					
7	Nota okumaya çalışmaktansa parçalarımı kulaktan dinleyerek çalmaya çalışırım.					
8	Sevdiğim bir eser olursa daha iyi çalışıp çabuk öğrenirim.					
9	Çalışırken metronom kullanmayı tercih ederim.					
10	Eserlerimi dizeklere bölerek çalışırım.					
11	Çalışacağım parçayı bir başka arkadaşımdan dinlemek beni güdülendirir.					
12	Hoca dersi sevdirirse öğrenci daha iyi çalışır ve öğrenir.					
13	Çalarken eserin cümlelerini bularak cümle çalışması yapmaya dikkat ederim.					
14	Her zaman parçalarımı cümle cümle çalışıp sonra birleştiririm.					
15	Notaları iyi okuyamadığım için ezber yapmayı tercih ederim.					
16	Çalışma isteğim eserin ezgisini sevmeme bağlıdır.					
17	Bir eserin nüanslarına mutlaka dikkat ederim.					
18	Yeni bir parça öğrenirken ölçü ölçü çalışırım.					
19	Kendime güvenerek çalışmam için parçamı bir başkasından dinlemem gerekir.					
20	Eğer dersi seversem her zaman daha hızlı öğrenirim.					
21	Çalışmaya başlamadan önce parmak egzersizi yaparım.					
22	Yeni parçalarımı her zaman dört ölçüye bölerek çalışmayı tercih ederim.					
23	Derse gitmeden önce kontrol amacı ile bir başka arkadaşıma parçamı çalarak fikrini alırım.					
24	Çalacağım parçayı inceleyip zorluk derecesini düşünürüm.					
25	Bir parça aldığımda hemen o parçanın armonik yapısını incelerim.					
26	Yeni bir parçayı öğrenmeye çalışırken sağ eli ayrı sol eli ayrı çalışmayı tercih ederim.					
27	Eserlerimi başkalarının çaldıklarını taklit ederek çıkarmaya çalışırım.					
28	Çalışmalarım bir eseri başından sonuna çok defa tekrar etmekle geçer.					
29	Deşifre yaparken parmak numarasına bakarak uygulamak benim için önemlidir.					
30	Her zaman notaları ezberlemeye çalışırım.					

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Use of Full Hierarchy Consistency Index to Assess Response Consistency

Lokman Akbay, Mustafa Kılınç

To cite this article: Akbay, L., & Kılınç, M. (2018). Use of Full Hierarchy Consistency Index to Assess Response Consistency. *International Journal of Assessment Tools in Education*, 5(1), 105-118. DOI: 10.21449/ijate.350499

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 105-118

http://www.ijate.net

e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Use of Full Hierarchy Consistency Index to Assess Response Consistency

Lokman Akbay[*], **Mustafa Kılınç**

Educational Measurement and Evaluation, Mehmet Akif Ersoy University, Burdur, Turkey

Abstract: Measurement models need to properly delineate the real aspect of examinees' response processes for measurement accuracy purposes. To avoid invalid inferences, fit of examinees' response data to the model is studied through *person-fit* statistics. Misfit between the examinee response data and measurement model may be due to invalid models and/or examinee's aberrant response behavior such as cheating, creative responding, and random responding. Hierarchy consistency index (HCI) was introduced as a person-fit statistics to assess classification reliability of particular cognitive diagnosis models. This study examines the HCI in terms of its usefulness under nonhierarchical attribute conditions and under different item types. Moreover, current form of HCI formulation only considers the information based on correct answers only. We argue and demonstrate that more information could be obtained by incorporating the information that may be obtained from incorrect responses. Therefore, this study considers the full-version of the HCI (i.e., FHCI). Results indicate that current form of HCI is sensitive to misfitting item types (i.e., basic or more complex) and examinee attribute patterns. In other words, HCI is affected by the attribute pattern an examinee has as well as by the item s/he aberrantly responded. Yet, FHCI is not severely affected by item types under any examinee attribute pattern.

ARTICLE HISTORY

Received: 01 October 2017

Revised: 07 November 2017

Accepted: 20 November 2017

KEYWORDS

Person-fit,

Attribute Hierarchy Index,

Cognitive Diagnosis,

1. INTRODUCTION

Measurement models must play an important role in test construction and result interpretation processes of educational assessments. As a recent measurement model, cognitive diagnosis modeling has drawn great attention on the grounds of incorporating cognitive psychology in testing practices. Cognitive diagnosis models (CDMs) are the statistical models used to identify the knowledge and skills students mastered or failed to master in a particular domain. To accomplish this, associations between the test items and the measured knowledge or skills must be predefined. These measured knowledge, skills, cognitive processes, and problem solving steps are referred to as *attributes* (de la Torre, 2009; de la Torre & Lee, 2010) and the matrix reflecting items-by-attributes association is called *Q-matrix* (Tatsuoka, 1983). For example, if an item requires the first two attributes out of three attributes measured by a test, q-vector of this item is specified as [110] in the Q-matrix. Here 1 stands for required

[*]Corresponding Author E-mail: lokmanakbay@gmail.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.350499**

attribute and 0 indicates not required attribute. This vector signifies the fact that examinees are expected to be mastered the first two attributes to reach correct answer.

Starting with the pioneering work of Tatsuoka (1983), various approaches integrating cognitive theory into psychometric practices have been proposed. The rule space methodology (RSM: Tatsuoka, 1983), attribute hierarchy method (AHM: Leighton, Gierl, & Hunka, 2004), deterministic input, noisy "and" gate (DINA: Junker & Sijtsma, 2001), and generalized-DINA (GDINA: de la Torre, 2011) are among the examples of CDMs. In general, based on the presence of absence of *K* measured attributes, at most 2^K latent classes can be formed by a CDM where *K* indicates the number of attributes to be measured. For instance, when a test developed for cognitively diagnosis assessment measures three attributes, CDM analysis classifies examinees into, at most, eight possible latent classes (i.e., {000}, {100}, {010}, {001}, {110}, {101}, {011}, {000}). When an examinee is classified in {100} latent group, his/her estimated attribute pattern becomes [100], which indicates that the examinee has mastered the first attribute and has not mastered the second and third. The ultimate purpose of CDMs is to provide feedback on students' strengths and weaknesses based on the attribute pattern, which could be helpful to modify teaching and learning activities.

To evaluate examinees' performance, CDMs establish the relations between examinees' response data and their mastery status of attributes within measured domain. Probability of an examinee's correct response to a test item is modeled as a function of item parameters and examinee's mastery of the attributes (Cui & Leighton, 2009). For example, the DINA model assumes that an examinee correctly responds to an item as long as the examinee has mastered all the required attributes required for that item. Thus, for one item, examinees are spread into two distinct groups (i.e., examinees who have mastered all required attributes for the item and examinees lacking at least one required attribute). This group-specific deterministic response can be defined by

$$\eta_{lj} = \prod_{k=1}^{K} \alpha_{lk}^{q_{jk}}$$

where, η_{lj} is deterministic response of group *l* by item *j* (i.e., 1 or 0); K indicates total number of attributes measured by the test; α_{lk} is the group *l*'s mastery status of attribute *k*; and q_{jk} is the k^{th} element in the q-vector of item *j*, which indicates whether or not attribute *k* is required for correct response of item *j*.

Item response function (IRF) of the DINA model has a probabilistic component, which allows possibility of *guessing* (i.e., responding correctly when not all attributes are mastered) and *slip* (i.e., giving an incorrect response when all required attributes are mastered). Given examinee *i*'s observed response to item *j* (i.e., X_{ij}), these two item parameters are denoted as $g_j = P(X_{ij} = 1|\eta_{ij} = 0)$ and $s_j = P(X_{ij} = 0|\eta_{ij} = 1)$ for guessing and slip parameters, respectively. Given the item parameters, the IRF of the DINA model is written as

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i) = P(X_{ij} = 1|\eta_{ij}) = g_j^{(1-\eta_{ij})}(1 - s_j)^{\eta_{ij}}$$

where $\boldsymbol{\alpha}_i$ is the attribute pattern of examinee *i*; η_{ij} is the expected response of examinee *i* to item *j*; X_{ij} is examinee *i*'s observed response to item *j*; and g_j and s_j are the guessing and slip parameters of item *j* (de la Torre, 2009). For further information on the estimation and classification of the DINA model, readers may refer to de la Torre (2009).

Measurement accuracy of examinees is directly related to appropriateness of measurement model, which need to properly delineate the real aspect of examinees' response

processes (Cui & Leighton, 2009). For instance, when attributes hold a hierarchical structure (i.e., some of the attributes are prerequisite to master others), not all 2^K latent classes are permissible. Therefore, examinees' response data should be analyzed accordingly. Thus, identification of the attributes, attribute structure, and attribute specifications in the Q-matrix must be precise. Otherwise, invalid inferences about examinees' knowledge states could be made. Furthermore, to avoid invalid inferences, fit of examinees' response data to the model is studied through 'person-fit' statistics. By means of person-fit statistics, examinees who are not being measured well by the test are identified (Cui & Leighton, 2009). Misfit between the examinee response data and measurement model may be due to invalid models and/or examinee's aberrant response behavior (e.g., cheating, creative responding, and random responding).

Cui and Leighton (2009) have introduced a person-fit index to assess classification reliability of specific cognitive diagnosis models (e.g., attribute hierarchy model [AHM: Leighton, Gierl, & Hunka, 2004]). This person-fit index is referred to as hierarchy consistency index (HCI) as it was also used by Cui (2007) to measure the accuracy of specified hierarchical structure of attributes in AHM. More information on the index is provided below.

1.1. Hierarchy consistency index (HCI)

Cui and Leighton (2009) introduced a person-fit statistic to detect misfit between item responses and the cognitive model. This fit statistic is called hierarchy consistency index (HCI) and ranges from -1.0 to 1.0. Statistics close to 1.0 indicate good fit between examinee responses and the model whereas statistics close to -1.0 indicate misfit. Definition of HCI is given in equation 1, which is borrowed from Cui and Leighton (2009), p 436. As it would be seen from the formula on Figure 1, HCI operates based on the match between an examinee's observed item responses and expected item responses based on a hierarchical relationships among measured attributes.

$$HCI_i = 1 - \frac{2\sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{ij}(1 - X_{ig})}{N_{C_i}}$$

where X_{ij} is examinee *i*'s binary response to item *j* where 0 indicates incorrect response and 1 stands for a correct response; $S_{correct_i}$ is an index set that includes items requiring the subset of attributes required by item *j* when examinee's response to item *j* is correct; X_{ig} is examinee *i*'s response to item g where item g belongs to $S_{correct_i}$; and N_{C_i} is the total number of comparisons for all the items correctly responded by examinee *i*.

2. ARGUMENT

When index is computed solely for the correct responses, some correct responses require less comparison than others. For example, imagine a test measuring three hierarchically structured attributes, in which attribute-1 (A1) is the most basic and attribute-3 (A3) is the most complex attribute. Here, when an item requiring A3 is correctly answered by an examinee, all other responses of the examinee are also expected to be correct. Thus, all other item responses are considered in index computation. Yet, when an examinee correctly responses an item requiring only A1 (i.e., the most basic attribute) only, only the items requiring sole A1 are considered for HCI computation. The potential problems in this regard are depicted below in a scenario where three hierarchical attributes are measured by a 10-items test, for which the Q-matrix is given in Table 1 and hierarchical structure of attribute is given in Figure 1.

Table 1. Q-matrix for 10-items test

Items	A1	A2	A3
1	1	0	0
2	1	1	0
3	1	1	1
4	1	0	0
5	1	1	0
6	1	1	1
7	1	0	0
8	1	1	0
9	1	1	1
10	1	0	0

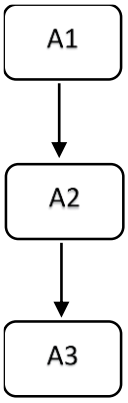


Figure 1. Linear hierarchy of three attributes

When an examinee's true attribute pattern is [000], expected responses of the examinees to all items becomes incorrect (i.e., 0). However, because of probabilistic component of the models, this examinee may correctly respond to one item. When we consider this *guessed* item only in HCI computation, all the comparisons we do will yield a misfit. Thus the computed HCI will be -1, which will, in turn, indicate that this examinee's responses do not fit to model. In fact, there is only one response that contradicts with the model expectancy. Imagine another examinee whose true attribute pattern is [111]. In this case expected responses of this examinee will be all correct. When the examinee misses one item, then only the comparisons due to that item will be left. Moreover, among the all comparisons conducted for the correct responses, only this incorrect response will yield misfit. There will be some reduction in the HCI due to this one misfit, yet the impact of this *slipped* item will not be as large as it is in previous case. Furthermost, because it will change the comparisons counted toward HCI, items missed by the examinee also matter.

Table 2. Two examinees and their HCI indices based on hypothetical response patterns

Examinees	Attribute profile	Response data	HCI
E1	000	**1**000000000	-1.000
E1	000	00**1**0000000	-1.000
E2	111	**0**111111111	0.667
E2	111	11**0**1111111	0.917

This scenario and resulting HCIs are summarized in Table 2. When E1 (i.e., an examinee with an attribute pattern [000]) guesses only one item, than HCI becomes -1. When E2 (i.e., an examinee with an attribute pattern [111]) slips one item, than HCI becomes smaller than 1.0, yet impact of slipped item is determined by the q-vector of the item. In other words, whether slipped item requires basic attribute or complex attribute matters. In above case, when an item requiring the most basic attribute is missed, HCI becomes .667. Impact of missed item when it requires the most complex attribute is relatively smaller (i.e., computed HCI is .917). As can be seen, although there is only one misfitted item in all cases, their impact on examinees' response consistency is different under different conditions.

2.1. Full Hierarchy Consistency Index

It should be noted here that *guessing* does not necessarily mean random guessing in cognitive diagnosis modelling framework, rather it means completing a task employing any other strategy that is not specified by the model. Therefore, guessing and slip behaviour of

examinees may be different for items requiring basic ore more complex attributes. From this point of view, consistency index should not be dramatically affected by the attribute-and-item specification of misfitted item. One possible way to control this is to consider all items for examinee response fit, which can be implemented by adding a second component to HCI that includes comparisons for item sets consists of items that are expected to be incorrectly responded by the examinee. Then the *full* version of the index may be represented as

$$FHCI_i = 1 - \frac{2[\sum_j \sum_{j' \in S_{j-correct}} X_{ij}(1 - X_{ij'}) + \sum_j \sum_{j'' \in S_{j-incorrect}} (1 - X_{ij}) X_{ij''}]}{N_{C_i}}$$

where X_{ij} is examinee i's binary response to item j where 0 indicates incorrect response and 1 stands for a correct response; $S_{j-correct}$ is an index set that includes items requiring the subset of attributes required by item j when examinee's response to item j is correct; $S_{j-incorrect}$ is an index set that includes items requiring all the attributes required by item j when the item incorrectly answered by the examinee; $X_{ij'}$ is examinee i's response to item j' where item j' belongs to $S_{j-correct}$; $X_{ij''}$ is examinee i's response to item j'' where item j'' belongs to $S_{j-incorrect}$; and N_{C_i} is the total number of comparisons for all the items responded by examinee i. This full version of the index will be referred to as *full hierarchy consistency index (FHCI)* throughout this paper. Computed FHCI indices for two previous examinees with certain response patterns are given in Table 3. Results based on FHCI are quite acceptable under all conditions.

Table 3. Two examinees and their FHCI indices based on the response patterns

Examinees	Attribute profile	Response data	HCI	FHCI
E1	000	**1**000000000	-1.000	0.765
E1	000	00**1**0000000	-1.000	0.438
E2	111	**0**111111111	0.667	0.429
E2	111	11**0**1111111	0.917	0.840

This study aims to focus on the following question:

- How successfully HCI is used under nonhierarchical attribute conditions (i.e., unstructured attribute cases) to identify aberrantly responded examinees,
- What is the impact of q-vector of a misfitting item on the HCI. More specifically, this study aims to unveil the impact of a misfitting item on HCI when it measures basic or more complex attributes,
- What is the distribution of misfitting examinees when number of misfits is equal across all permissible latent classes,
- Current form of HCI formulation only considers the information based on correctly answered items. Thus, more information could be obtained by incorporating the information that may be obtained from incorrect responses. Therefore, this study considers the Full-version of the HCI such that examinees' all responses rather than only correct responses are taken into account for consistency index computation.

3. METHOD

A simulation study and a real data analysis were conducted. In the simulation study, number of examinees, number of items and number of attributes were fixed to 2000, 20, and 6; respectively. Corresponding Q-matrix (i.e., item-by-attribute matrix) is given in Table 4. Corresponding Q-matrices for linear and divergent cases are given in Appendices. In the item response data generation, uniform examinee distribution was assumed. Two types of

hierarchical structures (i.e., linear and divergent) and an unstructured attribute case were considered. These hierarchical attribute structures can be seen in Figure 2. Four types of item misfits were considered:

Table 4. Generating Q-matrix

	Attributes							Attributes					
Items	A1	A2	A3	A4	A5	A6	Items	A1	A2	A3	A4	A5	A6
1	1	0	0	0	0	0	11	0	0	0	0	1	1
2	0	1	0	0	0	0	12	1	0	0	0	0	1
3	0	0	1	0	0	0	13	1	1	1	0	0	0
4	0	0	0	1	0	0	14	0	1	1	1	0	0
5	0	0	0	0	1	0	15	0	0	1	1	1	0
6	0	0	0	0	0	1	16	0	0	0	1	1	1
7	1	1	0	0	0	0	17	1	0	0	0	1	1
8	0	1	1	0	0	0	18	1	1	0	0	0	1
9	0	0	1	1	0	0	19	1	0	0	0	0	0
10	0	0	0	1	1	0	20	0	0	0	0	0	1

1. Creative responding (high guessing and slip in items requiring basic attributes)
2. Difficult (high slip in the complex items only)
3. Logical (high guessing in the items requiring basic attributes and high slip in the items requiring complex attributes)
4. Uniform (distribution of guessing and slip is uniform across all items)

For the *creative response* items, the lowest and highest success probabilities (i.e., P(0) and P(1)) were generated from U(0.20, 0.30) and U(0.70, 0.80), respectively, for items requiring basic attributes. These probabilities drawn from U(0.10, 0.20) and U(0.80, 0.90), respectively, for items requiring complex attributes. Lowest success probability of both basic and complex items in the *difficult* item case were generated from U(0.10, 0.20). In contrast, the highest success probabilities were generated from U(0.80, 0.90) and U(0.70, 0.80), respectively, for the basic and complex items. In the *logical* item case, the lowest and highest success probabilities were generated from U(0.20, 0.30) and U(0.80, 0.90), respectively, for items requiring basic attributes. Corresponding distributions for the complex item case were U(0.10, 0.20) and U(0.70, 0.80), respectively. Lastly, the lowest and highest success probabilities of examinees for both basic and complex items were generated from U(0.10, 0.20) and U(0.80, 0.90), respectively. These conditions are summarized in Table 5.

HCI and FHCI were employed to demonstrate extra information that can be obtained from incorrect responses. The data generation was based on the DINA model (de la Torre, 2009; Junker and Sijtsma, 2001). Throughout the study data generation performed using the OxMetrics programming language (Doornik, 2011) and index computation was performed in R-version 3.3.3. Simulation study is followed by a real data analysis. Data consist of 2922 examinees' binary responses to the 28 items in the grammar section of the ECPE examination. The test was developed and administered in University of Michigan English Language Institute in 2003. The dataset and the Q-matrix are available in and obtained from the `CDM' package (Robitzsch, Kiefer, George, & Uenlue, 2014) in R software environment.

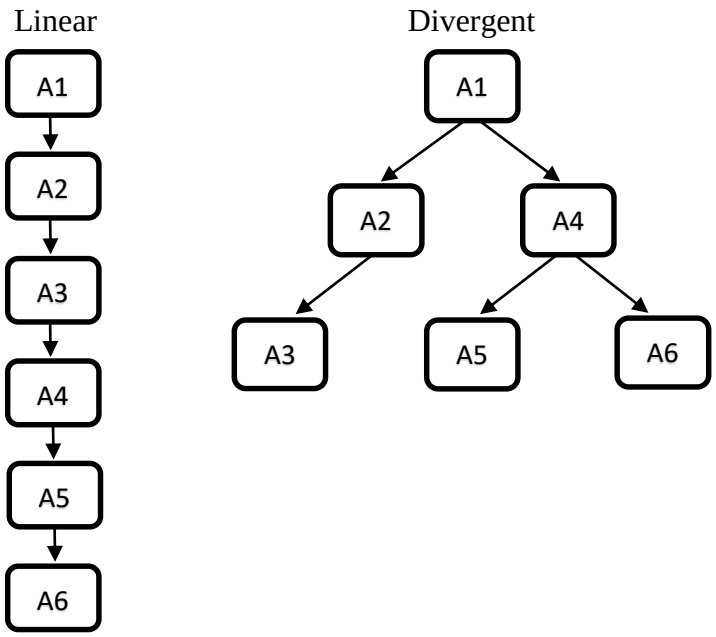


Figure 2. *Linear and divergent hierarchical structures.*

Table 5. Success probability distributions of item types

Item Types	Items with basic attributes		Items with complex attributes	
	P(0)	*P(1)*	*P(0)*	*P(1)*
Creative response	U(0.20, 0.30)	U(0.70, 0.80)	U(0.10, 0.20)	U(0.80, 0.90)
Difficult	U(0.10, 0.20)	U(0.80, 0.90)	U(0.10, 0.20)	U(0.70, 0.80)
Logical	U(0.20, 0.30)	U(0.80, 0.90)	U(0.10, 0.20)	U(0.70, 0.80)
Uniform	U(0.10, 0.20)	U(0.80, 0.90)	U(0.10, 0.20)	U(0.80, 0.90)

4. RESULTS

4.1. Simulation Results

Simulation results based on the HCI are given in Figure 3 as a matrix of scatterplots depicting HCI distribution of 2000 examinees where examinees are ordered based on the number of attributes they mastered. For instance, first a few hundreds of examinees in the linear case have the generating attribute pattern of [000000]; while very last a few hundreds have the generating attribute pattern of [111111]. Considering this order and the fact that all examinees' fit levels are approximately equal, it's very clear from the figure that HCI tends to be negative when an examinee has mastered smaller number of measured attributes. This reality emerges from the fact that when examinee guesses an item all other items requiring the subset of attributes specified in the guessed item are counted toward comparisons employed in index computation. HCI may be a good indicator of person fit when examinee has mastered most of the attributes, however, it may not be a good indicator for examinees who have lack of many attributes.

It can also be observed from Figure 3 that when number of latent classes decreases (i.e., hierarchy becomes more stringent) variance of HCI distribution shrinks. For example, in all types of item cases, HCI variance across attribute patterns is smaller when attributes are linearly structured. When attributes have no hierarchical structure (i.e., unstructured attribute case), HCI for examinees in any latent class are more disperse. Although item types do not make substantial differences, slight changes in the scatter plots by item types are observed. For

instance, in the difficult item case (i.e., high slips in the complex items only), HCI distribution of examinees who mastered more than half of the attributes are more disperse than the distribution of examinees who mastered a few attributes. Similarly, when creative item types are administered, variance of HCI of examinees lacking complex attributes elevates. These results are not surprising because when probabilistic component of item responses increases, examinees' observed responses deviate from the expected responses such that person-fit reduces.

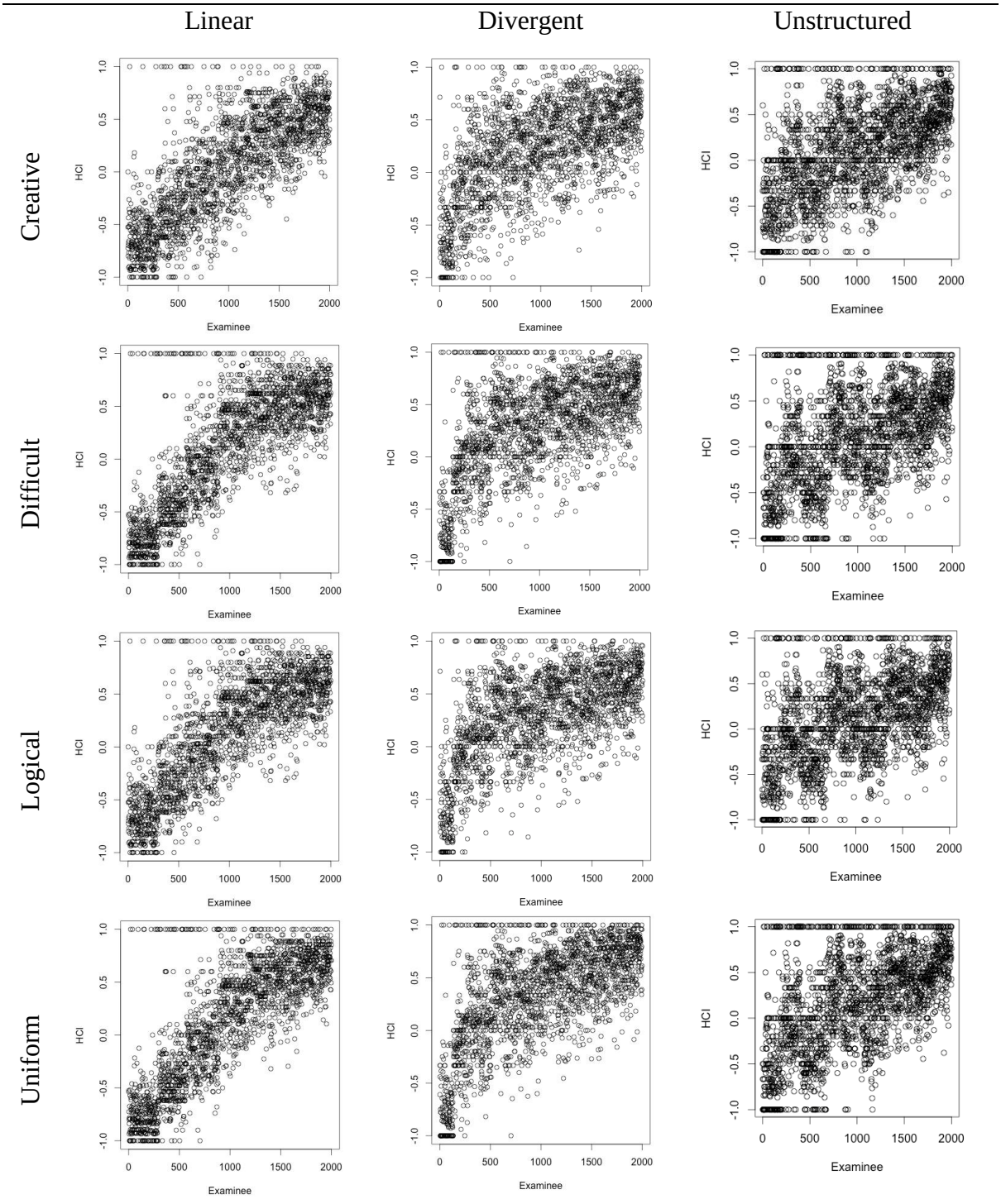


Figure 3. Matrix of scatterplots of HCI under various item types and attribute hierarchies

One major purpose of this study was to unveil the general improvement in identifying person-fit when not only correct responses but also incorrect responses are considered in person fit index computation. Results based on the FHCI are given in Figure 4. It can easily be seen at

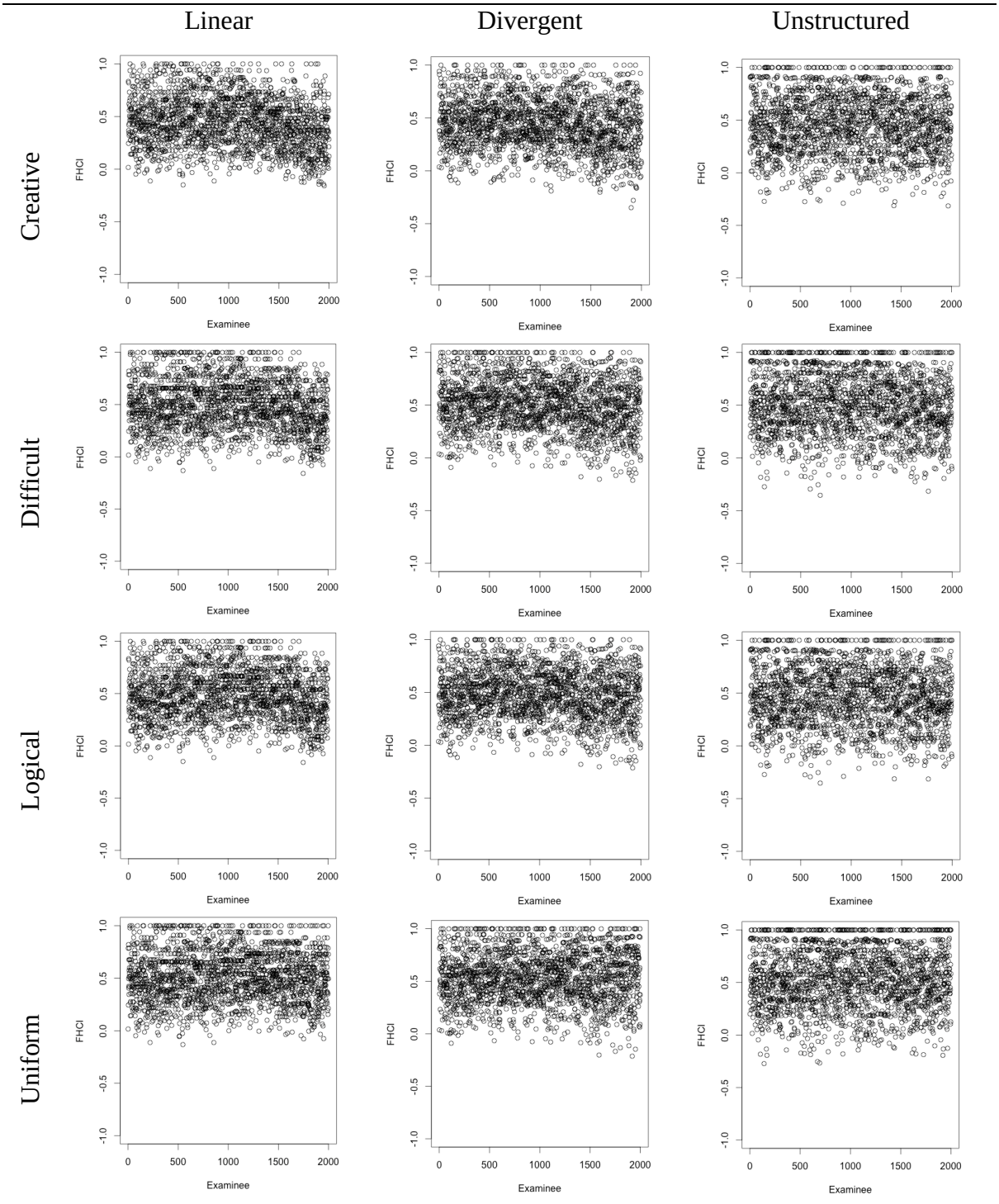


Figure 4. Matrix of scatterplots of FHCI under various item types and attribute hierarchies

first glance that, regardless of item type, attribute structure, and latent class an examinee is in, person-fit approximately falls between 0.00 and 1.00. This result suggests that FHCI may be considered as a more accurate person-fit index as it is not affected by examinees' attribute

pattern distribution (i.e., it measures fit in the same level of accuracy when examinee has mastered all or none of the measured attributes). Moreover, attribute structure does not significantly affect the results (i.e., variance of fit indices in the scatterplots are almost equal across linear, divergent, and unstructured attribute cases). Lastly, when FHCI is employed, small differences arising out of item types (i.e., creative, difficult, logical, and uniform) also diminished or even disappeared.

4.2. Real Data Analysis

Binary responses of 1922 examinees to 28 grammar items in the examination for the certificate of proficiency in English (ECPE) examination were analyzed in terms of examinees' person-fits. Q-matrix of the test and the data were obtained from 'CDM' package in R software environment. The data were analyzed previously by Templin and Bradshaw (2014) and specified a linear hierarchy among the three attributes (i.e., lexical rules, cohesive rules, and morphosyntactic rules) test is measuring. Scatter plots of examinees' person-fit results obtained by employment of HCI and FHCI are given Figure 5. When we look at the figure, FHCI result consistent with the simulation results, while HCI shows relatively better person-fit than what was observed in the simulation results.

However, remember that HCI fails to detect true person-fit when examinees did not master measured attributes. Assuming that the test truly measured aforementioned attributes and Q-matrix is correctly specified, correct answer proportions (proportion-corrects) of items may reflect attribute-pattern distribution of examinees. Proportion-correct of items are given in Table 6. Minimum and maximum proportion-corrects are .45 and .90, respectively. Moreover, 19 out of 28 items have been correctly answered by and over 70% of examinees, while only three items have been correctly answered by less than 50% of examinees. These results imply that many examinees in the sample have mastered two to three attributes. In the light of above information, person-fit result based on HCI could be more reflective of simulation results if there were more examinees lacking more than half of the attributes in the sample.

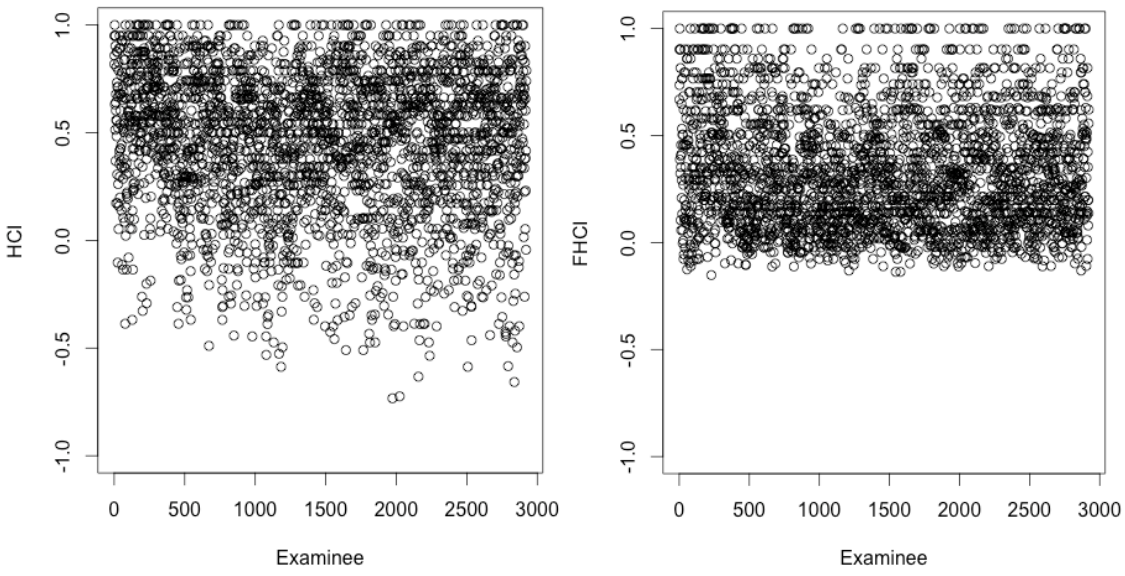


Figure 5. Scatter plots obtained by HCI and FHCI for ECPE data

Table 6. Proportion correct

Items	Proportion correct	Items	Proportion correct	Items	Proportion correct	Items	Proportion correct
1	.80	8	.90	15	.88	22	.63
2	.83	9	.70	16	.70	23	.81
3	.58	10	.66	17	.89	24	.53
4	.71	11	.72	18	.85	25	.62
5	.89	12	.43	19	.71	26	.70
6	.85	13	.75	20	.46	27	.45
7	.72	14	.65	21	.76	28	.82

min.=.43; mean=.71; max.=.90

5. CONCLUSION

HCI and FHCI have been employed under various conditions in this research. In data generation procedure guessing and slip for any item types did not exceed .30 (i.e., maximum P(0) = U(.20, .30) and minimum P(1) = U(.70, .80)). Thus, all examinees with different attribute patterns fit to the model equally well. Results suggested that HCI is a good indicator of person-fit as long as examinee has mastered most of the attributes. However, it fails to capture fitting examinees when examinees lack of many attributes. Conversely, FHCI may be considered as a more accurate person-fit index as it is not affected by examinees' attribute pattern distribution (i.e., it measures fit in the same level of accuracy when examinee has mastered all or none of the measured attributes).

Furthermore, FHCI is robust to different types of items such that impacts of misfit on basic and complex items are comparable. Therefore, more correct results yielding accurate inferences may be obtained by employment of FHCI. Study results demonstrated that regardless of item type, attribute structure, and latent class an examinee is in, FHCI approximately falls between 0.00 and 1.00. These results may be considered to form a cut-off to make a decision when FHCI is used to determine whether an examinee's responses fit to model. So, as long as an examinee's FHCI is positive (i.e., larger than .00), we may postulate this person's fit to model as acceptable. Lastly, in cases where we use FHCI as a measure of hierarchy consistency (i.e., whether assumed hierarchy for the model is acceptable), we should look for the distribution of examinees' FHCI, which need to be ranging from .00 to 1.00.

Disclosure Statement

No potential conflict of interest was reported by the authors.

6. REFERENCES

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46,* 429-449.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34,* 115-130.

Doornik, J. A. (2011). *Object-oriented matrix programming using Ox* (Version 6.20). London: Timberlake Consultants Press.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258-272.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuokas rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187-212.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2014). *CDM: Cognitive diagnosis modeling*. R package version 5.8-9. https://CRAN.R-project.org/package=CDM

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79,* 317-339.

7. APPENDICES

Appendix A. Q_matrix by the linear attribute structure

	Attributes							*Attributes*					
Items	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*	*Items*	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*
1	1	0	0	0	0	0	*11*	1	1	1	1	1	1
2	1	1	0	0	0	0	*12*	1	1	1	1	1	1
3	1	1	1	0	0	0	*13*	1	1	1	0	0	0
4	1	1	1	1	0	0	*14*	1	1	1	1	0	0
5	1	1	1	1	1	0	*15*	1	1	1	1	1	0
6	1	1	1	1	1	1	*16*	1	1	1	1	1	1
7	1	1	0	0	0	0	*17*	1	1	1	1	1	1
8	1	1	1	0	0	0	*18*	1	1	1	1	1	1
9	1	1	1	1	0	0	*19*	1	0	0	0	0	0
10	1	1	1	1	1	0	*20*	1	1	1	1	1	1

Appendix B. Q_matrix by the divergent attribute structure

	Attributes							*Attributes*					
Items	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*	*Items*	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*
1	1	0	0	0	0	0	*11*	1	0	0	1	1	1
2	1	1	0	0	0	0	*12*	1	0	0	1	0	1
3	1	1	1	0	0	0	*13*	1	1	1	0	0	0
4	1	0	0	1	0	0	*14*	1	1	1	1	0	0
5	1	0	0	1	1	0	*15*	1	1	1	1	1	0
6	1	0	0	1	0	1	*16*	1	0	0	1	1	1
7	1	1	0	0	0	0	*17*	1	0	0	1	1	1
8	1	1	1	0	0	0	*18*	1	1	0	1	0	1
9	1	1	1	1	0	0	*19*	1	0	0	0	0	0
10	1	0	0	1	1	0	*20*	1	0	0	1	0	1

Appendix C. R Scripts written to compute HCI and FHCI

```
########### HCI #########
setwd("~/Desktop/FHCI/data")
data<-read.table("ResponseData.txt", header=F, sep="")
q<-read.table("Q_matrix.txt", header=F, sep="")
p<-matrix(NA,1,nrow(data)) # person sayisisi kadar (samplesize)
for(i in 1:nrow(data)){
J=nrow(q)
m<-matrix(NA,1,nrow(q)) # misfit: madde sayisisi kadar
nci<-matrix(NA,1,nrow(q)) # total number of comparison: madde sayisisi kadar
for(j in 1:J){
   c<-matrix(NA,1,nrow(q)) # comparison for item j
   for(l in 1:J){
   c[,l]<-ifelse(data[i,j]==1,(ifelse(sum(ifelse(q[j,]>=q[l,],1,0))==ncol(q),1,0)),0)}
   cj<- (sum(c)-(sum(data[i,]*c))) # number of misfit by item j
   m[,j]<-ifelse(data[i,j]==1,cj,0)
   nci[,j]<-sum(c)  # item j is compared with itselft, which should not be counted}
   HCIi<-1-(2*(sum(m)/(sum(nci)-sum(data[i,])+.000001)))  # .0001 is to avaid NaN result for 0
response vectors
   p[,i]<-HCIi}
plot(p[1,], xlab="Examinee", ylab="HCI")
```

```
########### FHCI #########
setwd("~/Desktop/FHCI/data")
data<-read.table("ResponseData.txt", header=F, sep="")
q<-read.table("Q_matrix.txt", header=F, sep="")
data1<-matrix(NA,nrow(data),nrow(q))
for(i in 1:nrow(data)) {
  for(j in 1:nrow(q)){
  data1[i,j]<-ifelse(data[i,j]==0,1,0)}}
 p<-matrix(NA,1,nrow(data)) # person sayisisi kadar (samplesize)
 for(i in 1:nrow(data)){
 J=nrow(q)
 m<-matrix(NA,1,nrow(q)) # misfit: madde sayisisi kadar
 nci<-matrix(NA,1,nrow(q)) # total number of comparison: madde sayisisi kadar
 m1<-matrix(NA,1,nrow(q)) # misfit: madde sayisisi kadar
 nci1<-matrix(NA,1,nrow(q)) # total number of comparison: madde sayisisi kadar
 for(j in 1:J){
   c<-matrix(NA,1,nrow(q)) # comparison for item j
   c1<-matrix(NA,1,nrow(q)) # comparison for item j
   for(l in 1:J){
   c[,l]<-ifelse(data[i,j]==1,(ifelse(sum(ifelse(q[j,]>=q[l,],1,0))==ncol(q),1,0)),0)
   c1[,l]<-ifelse(data1[i,j]==1,(ifelse(sum(ifelse(q[j,]<=q[l,],1,0))==ncol(q),1,0)),0)}
   cj<- (sum(c)-(   sum(data[i,]*c))) # number of misfit by item j
   m[,j]<-ifelse(data[i,j]==1,cj,0)
   nci[,j]<-sum(c)  # item j is compared with itselft, which should not be counted
   cj1<- (sum(c1)-(sum(data1[i,]*c1))) # number of misfit by item j
   m1[,j]<-ifelse(data1[i,j]==1,cj1,0)
   nci1[,j]<-sum(c1)  # item j is compared with itselft, which should not be counted}
HCIi<-1-(2*((sum(m)+sum(m1))/(sum(nci)-sum(data[i,])+sum(nci1)-sum(data1[i,])+.000001)))  #
.0001 is to avaid NaN result for 0 response vectors
p[,i]<-HCIi}
plot(p[1,], xlab="Examinee", ylab="FHCI", ylim=c(-1,1))
```

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Investigation of 9th Grade High School Students' Attitudes towards Science Course

Orhan Karamustafaoğlu, Adem Bayar

To cite this article: Karamustafaoğlu, O., & Bayar, A. (2018). Investigation of 9th Grade High School Students' Attitudes towards Science Course. *International Journal of Assessment Tools in Education*, 5(1), 119-129. DOI: 10.21449/ijate.365073

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 119-129

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Investigation of 9th Grade High School Students' Attitudes towards Science Course

Orhan Karamustafaoğlu*, Adem Bayar

Amasya University, Faculty of Education, Amasya, Turkey

Abstract: In this study, ninth grade students' attitudes towards science were investigated in terms of self-regulation strategies, motivational beliefs and gender variables. The sample of this study includes 322 male and 296 female in total 618 students from 3 different high schools (Science high school, Anatolian high school, and Vocational high school) in center district of Amasya city. To collect the data, the researchers employed "Motivated Strategies for Learning Questionnaire" which has been developed by Pintrich and De Groot in 1990, adapted into Turkish by Uredi in 2005 and consists of 44 items and "Colorado Learning Attitudes about Science Survey (CLASS)" has been developed by Adams and others in 2006, adapted into Turkish by Bayar and Karamustafaoğlu in 2015 and consists of 36 items. For data analysis, mean, standard deviation, independent t-test and correlation were addressed. The results of this study show that there are statistically significant relationships between 9th grade students' attitudes towards science and self-regulation strategies, motivational beliefs, and gender.

ARTICLE HISTORY

Received: 22 August 2017
Revised: 05 December 2017
Accepted: 14 December 2017

KEYWORDS

Attitudes towards science,
Gender,
Self-regulation,
Motivation

1. INTRODUCTION

Knowing the whole character of learners including psychological, cognitive, social, and emotional development is very important for educators to access the success in education. Due to the fact that the research studies in education place so much emphasis on cognitive development of learners and ignore other development levels including emotional development (Akbaş, 2004; Selvi, 1996). Educators consider the cognitive learning as the basement for their instruction and disregard emotional learning (Bacanlı, 1999; Bilen, 2001). Especially, by the beginning of 2000s, it is understood that intellectual factors are not solely enough for students' learning and academic achievement. In addition to intellectual factors, it has been accepted that emotional factors are important for students' learning and academic achievement. Since, emotional learning is the tool for cognitive learning (Demirbaş & Yağbasan, 2004). At this point, it is worth mentioning various definitions of "attitude" which is one of the aspects of the emotional learning. Attitude is identified as the tendency consisting of both negative and positive behaviours towards any object, events, or people (Eagly & Chaiken, 1993; Petty, 1995; Turgut, 1997). In Ulgen's (1997) points of view, the attitude is a phenomenon that influences

*Corresponding Author E-mail: orseka@yahoo.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.365073**

individuals' decision-making process. On the other hand, some intellectuals (Ajzen & Fishbein, 1980; Safran, 1993) discuss the attitude with its cognitive, sensual, and behavioral characters. Individuals' interests and attitudes are important aspects for identifying the characters of individuals. Knowing their interest and aspects might help anticipate future actions of individuals (Tekin, 1996).

In science education, attitude is the tendency to evaluate facts, events, and objects related to science (Gardner, 1975). Examining the body of research in science education about attitudes shows that attitudes of learners towards science education and their scientific attitudes are widely elaborated by researchers (Byrne & Johnstone, 1988; Koballa, 1988). While cognitive factors are emphasized in scientific attitudes, sensual aspects are paid more attentions in attitudes towards science education (Hamurcu, 2002). The results of the studies exploring learners' attitudes towards science education and physics education are very similar to each other. Even though physics is a relevant and applicable area to the everyday life, students think physics is very boring and challenging (Sarı, 2015; Tekbıyık & Akdeniz, 2010). The main prerequisite to teach physics effectively is to take attention of students to science or physics and suggest them to alternative learning strategies (Whitelegg & Parry, 1999; Zacharia, 2003). In a study, Ulgen (1997) explores students' beliefs about physics education. Participants' responses to the question that "Do you think physics course must be compulsory in high schools?" varies and as following: **a)** students fail physics, therefore it must not be taught in schools **b)** I hate physics **c)** Physics is a beneficial class, I apply what I learn in physics class to everyday life**. d)** I like doing physic homework **e)** Learning physics is essential for everybody, **f)** It does not matter for me whether physics is taught in schools or not. Among students' responses, b represents students' negative feelings, d represents positive feelings of students and c represents students' beliefs on cognitive side of physics. In physics classes, the questions like "Why do I have to learn all these facts" or "Where and when would I use what I learn in Physics class" come to students' minds. To prevent negative feelings and attitudes towards science and physics, it is necessary to teach children why science is important and a requirement for their lives. In the new science education program updated in 2013, the importance of beliefs and attitudes towards science education is emphasized in order to attract students to science classes (MONE, 2013). One of the purposes of the new science education program is to help students develop positive attitudes towards science. It is discussed that showing children the relationship among science, technology, environment, and community might increase students' interests in science. It is also argued that new research should be conducted to explore the factors influencing students' attitudes.

In the relevant literacy in science education, the issue regarding whether gender influence students' attitudes towards science is widely discussed. While some studies (Demirci, 2004; Güngör & Eryılmaz, 2006; Osborne, Simon & Collins, 2003) consider gender as one of the important factors influencing students' attitudes towards science, other studies make opposite arguments (e.g Barrington & Hendicks, 1988; Çakır, Şenler & Taşkın, 2007; Sorge, 2007). Şengören, Tanel and Kavcar (2006) found in their study that the students' attitudes towards optics, that is one branch of physics or science, were not changed by gender differences. In another study, Çakır, Şenler and Taşkın (2007) have found that there is no relationship between students' attitudes towards science and gender. To make a further investigation about students' attitudes towards science education, various scales were developed. However, these scales are mostly developed to assess attitudes of secondary or university students. (Bilgin, Özarslan & Bahar, 2006; Bozdoğan & Yalçın, 2005; House & Prison, 1998; Kind, James & Barmby, 2007; Nuhoğlu, 2008; Nuhoğlu & Yalçın, 2004; Pell & Jarvis, 2001; Reid & Skrybina, 2002). Limited number of scales in literature is available to evaluate high school students' attitudes in science classes.

Self-regulation and motivation are two important emotional factors that affect learning. When the relevant literature examined different definitions of self-regulation have been found. For the first time, the concept of self-regulation has been raised by Albert Bandura in 1986. Accordingly, the meaning of self-regulation is that a person has active role on his or her learning and examines the status of teaching-learning process. Çiltaş (2011) described self-regulation as "to determine your own personal learning aims and in accordance with its principles to motivate yourself cognitively (p:3)". Pintrich (2000) defined that the concept of self-regulation is describing own personal learning purposes and actively participating in teaching-learning process in order to achieve these described purposes. In a similar vein, Kauffman (2004) identified the concept of self-regulation as editing students' different learning activities, controlling and managing all situations. In this regard, individuals are responsible for their own learning and arrange learning-teaching activities based on their own needs. This creates a new learning approach, which is named as self-regulated learning. According to that, self-regulated learning approach provides active participation opportunities for learners and meets with individuals' needs. As a result, self-regulated learning approach can be described as learning process that motivates individuals for learning (Altun & Erden, 2006). Çiltaş (2011) identified self-regulated learning approach as "the way of knowing yourself and all processes, techniques, tactics, and strategies that can be used for personal learning" (p:3). As understood from above statements, the keyword of self-regulation means the learners actively involve teaching and learning process.

The aim of this study was to examine ninth grade students' attitudes towards science in terms of self-regulation strategies, motivational beliefs and gender variables. In order to reach the aim of this study, the researchers have addressed the following research questions:

1) What are the level of 9th grade students' self-regulation strategies and motivational beliefs with attitudes towards science?
2) Do the 9th grade students' self-regulation strategies and motivational beliefs with students' attitudes towards science change by gender?
3) Is there any relationship between 9th students' self-regulation strategies, motivational beliefs and students' attitudes toward science?

2. METHOD

2.1. Research Model

In this descriptive study, survey model has been used. According to Karasar (2009), survey has been identified as research approaches that aim to describe past or present situation as it is or was. In this study, the relationship between a dependent variable (students' attitudes towards science) and independent variables (students' self-regulation strategies, motivational beliefs, and gender) has been tested.

Within the scope of survey model, "Motivated Strategies for Learning Questionnaire" and "Colorado Learning Attitudes about Science Survey (CLASS)" have been applied to collect data. Motivated Strategies for Learning Questionnaire has been developed by Pintrich & De Groot in 1990. Its original form was translated into Turkish by Uredi in 2005 and consists of 44 items. Colorado Learning Attitudes about Science Survey (CLASS) has been developed by Adams and others in 2006. Its original form was translated into Turkish by Bayar and Karamustafaoğlu in 2015 and consists of 36 items.

2.2. Study Group

This study has been conducted in the city center of Amasya, Turkey recruiting participants in different type of high schools (Science High School, Anatolian High School, and Vocational High School) governed by Ministry of National Education. Of 618 high school

students selected by a convenience sampling participated in the study, 96 are in Science High School, 277 are in Anatolian High School, and 245 are in Vocational High School. The sample of the study is consisting of 322 (52.1%) male and 296 (47,9%) female students who are taking physics course in 9th grade at these schools. The distributions of students according to high school types and gender have been shown in Table 1.

Table 1. Distributions of students according to gender and high school types

High School Types	Male	Female	Total
Science High School	42	54	96
Anatolian High School	126	151	277
Vocational High School	154	91	245
Total	322	296	618

2.2. Data Collection Tools

To collect data, three different data collection tools have been used for this study. These are: 1. Personal Information Form, 2. Motivated Strategies for Learning Questionnaire, and 3. Colorado Learning Attitudes about Science Survey (CLASS).

Personal Information Form

The researchers have used "Personal Information Form" to collect data regarding students' gender and high school type variables.

Motivated Strategies for Learning Questionnaire

The data has been collected applying Motivated Strategies for Learning Questionnaire, adapted to Turkish by Uredi in 2005 from its original form developed by Pintrich and De Groot in 1990. The Turkish version of scale, which has been employed to collect data in this study, consists of 44 items and uses a 7-point Likert-type scale. The 1-3 interval represents low level, 3-5 average level and, 5-7 high level.

This questionnaire consists of two sub-factors which are self-regulation strategies and motivational beliefs. These sub-factors separately have 22 items. By the process of adopting to Turkish version of this questionnaire, Cronbach's Alpha coefficient values have been calculated between .81 and .92 (Üredi, 2005). Similarly, the researchers have calculated the overall Cronbach's Alpha coefficient value and found it as .87, indicating strong internal consistency. Furthermore, the researchers have computed each sub-factors' internal consistency coefficient values and found them as .83 and .88 respectively.

Colorado Learning Attitudes about Science Survey (CLASS)

The data has been collected applying Colorado Learning Attitudes about Science Survey (CLASS), adapted to Turkish by Bayar and Karamustafaoğlu in 2015 from its original form developed by Adams et al., in 2006. They employed test-retest technique for the reliability of scale and found that the scale is reliable. Also, the Turkish form of the scale's internal consistency reliability has been found between .72 and .84 for each sub-factors and test-retest reliability coefficients varied between .85-.93 (Bayar & Karamustafaoğlu, 2015). The Turkish version of scale, which has been used to collect data in this study, consists of 36 items. The researchers have calculated the overall Cronbach's Alpha coefficient value and found it as 0.91, indicating strong internal consistency.

The scale consists of 8 sub-factors. The researchers have also analyzed each subfactor's internal consistency coefficients value and found it as .83 for Real world connection subscale,

as .78 for Personal interest subscale, as .80 for Sense making/effort subscale, as. 88 for Conceptual connections subscale, as .81 for Applied conceptual understanding, as .82 for Problem solving general subscale, as .83 for Problem solving confidence subscale, as .76 for Problem solving sophistication subscale. These findings clearly show that each sub-factor has strong internal consistency.

2.2. Data Analysis

The data has been collected by paper-based of *Colorado Learning Attitudes about Science Survey (CLASS) and Motivated Strategies for Learning Questionnaire* with 618 students in 9th grades at three different high schools. On the day of each survey administration, the researchers have personally visited each participating school and individually administered the paper-based survey utilizing "group administration" techniques during the school day. Once data collection was completed, the collected data had been analyzed using SPSS Version 21.0 statistical software.

In the process of data analyzing, first, the descriptive statistics such as mean and standard deviation has been examined. Then, in order to determine whether there is any relationship between gender, self-regulation strategies and motivational beliefs with students' attitudes towards science, independent *t* test has been addressed. Furthermore, in order to find the relationship between self-regulation strategies and motivational beliefs with students' attitudes towards science correlation has been applied. The researchers have considered the *p* value level of 0.05.

3. FINDINGS

The purpose of this current study is to examine whether there is any relationship between 9th grade students' attitudes towards science with self-regulation strategies, motivational beliefs, and gender variables. In this regard, to answer the first research question of this study "What are the level of 9th grade students' self-regulation strategies and motivational beliefs with attitudes towards science?", the average of each item and standard deviation have been calculated and shown in Table 2.

Table 2. The Scores of Students on Motivated Strategies for Learning Questionnaire and Colorado Learning Attitudes about Science Survey (CLASS)

Survey	n	X_{avg}	Min.	Max.	sd
Self-regulation strategies	618	4,75	1,00	6,89	0,83
Motivational beliefs	618	4,01	1,45	6,26	0,72
Attitudes towards science learning	618	3,16	1,87	5,24	0,68

As seen in Table 2, the average of 9th grade students on self-regulation strategies (x_{avg} = 4,75, sd = 0,83) is higher than the average of 9th grade students on motivational beliefs (x_{avg} = 4,01, sd = 0,72). Moreover, 9th grade students' self-regulation strategies and motivational beliefs scores are on average. As shown in Table 2, 9th grade students' attitudes towards science scores (x_{avg} = 3,16, sd = 0,68) are on average. In the light of these results, it can be said that 9th grade students' self-regulation strategies, motivational beliefs, and attitudes towards science learning are on average.

Furthermore, the second research question of this study has been asked for serving the purpose of study. In order to answer the second research question of this study, "Does the 9th grade students' self-regulation strategies and motivational beliefs with students' attitudes towards science change by gender?", the data has been analyzed and the findings of t-test have been shown in Table 3.

Table 3. The Independent t-test Results of 9th Grade Students on Motivated Strategies for Learning Questionnaire and Colorado Learning Attitudes about Science Survey (CLASS) by Gender

Survey	Gender	n	X_{avg}	sd	t	p
Self-regulation strategies	Male	322	4,55	0,86	2,185	0.00*
	Female	296	4,96	0,80		
Motivational beliefs	Male	322	3,90	0,75	1,981	0.01*
	Female	296	4,13	0,71		
Attitudes towards science learning	Male	322	3,03	0,70	1,976	0.01*
	Female	296	3,33	0,66		

As seen in Table 3, the difference about the average of 9th grade male and female students on self-regulation strategies has been compared by t-test (t= 2,185; p<.05) and found it as statistically significant in favor of female students. The difference concerning the average of 9th grade male and female students on motivational strategies has been compared by t-test (t= 1,981; p<.05) and found it as statistically significant in favor of female students. Furthermore, the difference regarding the average of 9th grade male and female students' attitudes towards science learning has been compared by t-test (t= 1,976; p<.05) and found it as statistically significant in favor of female students.

The results of correlation analysis explaining the relationship of 9th grade students' scores on *Motivated Strategies for Learning Questionnaire and* Colorado Learning Attitudes about Science Survey (CLASS) have been shown in Table 4.

Table 4. The results of correlation analysis explaining students' scores on *Motivated Strategies for Learning Questionnaire and* Colorado Learning Attitudes about Science Survey (CLASS)

Survey	n	r	p
Self-regulation strategies - Motivational beliefs	618	0,59	.000
Self-regulation strategies - Attitudes towards science learning	618	0.49	.000
Motivational beliefs - Attitudes towards science learning	618	0,42	.000

r: correlation coefficient, p<.05

As seen in Table 4, there is moderately relationship on 9th grade students' self-regulation strategies and motivational beliefs (r=0,59, p<.05). In a similar vein, there is moderately relationship on 9th grade students' self-regulation strategies and attitudes towards science learning (r=0,49, p<.05). Moreover, there is moderately relationship on 9th grade students' motivational beliefs and attitudes towards science learning (r=0,42, p<.05). When the correlation coefficients examined, it can be clearly seen that there is a positive and linear relationship on 9th grade students' scores on *Motivated Strategies for Learning Questionnaire and* Colorado Learning Attitudes about Science Survey (CLASS).

4. DISCUSSION, CONCLUSION AND SUGGESTIONS

This study has been conducted in Amasya, Turkey recruiting participants, 9th grade high school students, in three different types of high schools. In this study, the researchers have examined students' attitudes towards science course by comparing self-regulation strategies, motivational beliefs and gender variables. It has been seen that there is a statistically significant

relationship students' attitudes towards science course, self-regulation strategies, motivational beliefs and being male or female.

The results of this study about self-regulation strategies and motivational beliefs indicate that there is a statistically significant relationship between students' attitudes towards science with self-regulation strategies and motivational beliefs. When the related literature has been scrutinized, there are some studies that emphasize the importance of relationship between students' attitudes towards science and physics with self-regulation strategies and motivational beliefs (Demir, Öztürk & Dökme, 2012; Mujtaba, & Reiss, 2013; Pendergast, Lieberman-Betz & Vail, 2017; Reid & Skryabina, 2002; Uzun & Keleş, 2012; Yamaç, 2011; Yaman & Dede, 2007; Yenice, Saydam & Telli, 2012; Zhang, Ding, & Mazur, 2017).

Furthermore, the results of this study related to gender variable show that there are differences between male and female students' scores on Colorado Learning Attitudes about Science Survey (CLASS). In literature, while some of the previous studies support the findings of this study that have determined the relationship between student attitudes towards science/physics and gender (Hançer, 2008; Lowery, Bowyer & Padilla, 1980; Özyürek & Eryılmaz, 2001), other studies have opposite argument (Kaya & Böyük, 2011; Murphy & Whitelegg, 2006; Yeşilyurt, 2004). The potential reason for these differences can be expressed that students might have different attitudes towards different branches of science/physics (Şengören, Tanel & Kavcar, 2006).

Considering the findings of this study, one of the important tasks of science teachers is to take students' attention to science. They should explain students the importance of science and its contribution to students' daily lives. It is essential to announce students that science is not only required for exams, rather science is a part of life and knowing science facilitates individuals' daily action and behaviors. Also, teachers should tell students that science is a necessary course for everybody not only for students in science-mathematics education. Science teachers should explain students that science is not just consisting of complicated formulas, conversely, it is helpful for everyone to understand how world works.

This study is conducted considering self-regulation strategies, motivational beliefs and gender as variables. Different variables- such as high school types, subject area, different grades, and age- might be used for future studies. Moreover, the target population and sample of this study might be thought as limited. To take away this thought, a further study might be done in multiple cities with more diverse participants.

5. REFERENCES

Adams, W.K., Perkins, K.K., Podolefsky, N.S., Dubson, M., Finkelstein, N.D. & Wieman, C.E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, *Physical Review Special Topics - Physics Education Research,* 2, 010101.

Ajzen, I. & Fisbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior*. New Jersey: Englewood Cliffs, Prentice-Hall.

Akbaş, O. (2004). *Türk milli eğitim sisteminin duyuşsal amaçlarının ilköğretim II. kademedeki gerçekleşme derecesinin değerlendirilmesi [Evaluation of the degree of reaching of affective goals at the elementery level in Turkish national education system]*. Yayınlanmamış doktora tezi [Unpublished doctorate dissertation], Gazi Üniversitesi [Gazi University], Eğitim Bilimleri Enstitüsü [Institute of Educational Science], Ankara.

Altun, S. & Erden, M. (2006). Öğrenmede motive edici stratejiler ölçeğinin geçerlik ve güvenirlik çalışması [Validity and reliability study of the motivated strategies for learning questionnaire. *Edu7, 2*(1), 1-16.

Bacanlı, H. (1999). *Duyuşsal Davranış Eğitimi.* Ankara: Nobel Yayın Dağıtım.

Bandura, A. (1986). *Social Foundations of thought and Action*. New Jersey: Prentice-Hall, Inc.

Barrington, B.L. & Hendricks, B. (1988). Attitudes toward science and science knowledge of intellectually gifted and average students in third, seventh, and eleventh grades. *Journal of Research in Science Teaching, 25*, 679-687.

Bayar, A. & Karamustafaoğlu, O. (2015). The Colorado Learning Attitudes about Science Survey (CLASS): The study of validity and reliability, *International Journal of Assessment Tools in Education, 2*(1), 40-57.

Bilen, M. (2001). Kurumlarda İnsan İlişkilerinin Başarıya Etkisi. *2000 Yılında Türk Eğitim Örgütü ve Yönetimi Ulusal Sempozyumu*, Ankara.

Bilgin, G., Özarslan, M., & Bahar, M., (2006). Comparison of the Attitude to Science Lesson and Success on the Nature of Matter of the Primary 8th grade Field Dependent and Independent Students' Cognitive Students. *VII. National Science and Mathematics Education Congress*, Ankara-Turkey.

Bozdoğan, A.E, & Yalçın, N, (2005). Attitudes of the basic education school students grade 6, 7 and 8 towards subjects of the physics in the science courses. *Gazi University Journal of Kırşehir Education Faculty, 6*(1), 241-247.

Byrne, M.S. & Johnstone, A.H. (1988). Critical thinking and science education. *Studies in Higher Education, 25*(8), 325.

Çakır, K.N., Şenler, B. & Taşkın, G.B. (2007). İlköğretim II. kademe öğrencilerinin fen bilgisi dersine yönelik tutumlarının belirlenmesi [Determining the attitudes towards science course of second grade students in primary school]. *The Journal of Turkish Educational Sciences, 5*(4), 637-655.

Çiltaş, A. (2011). Eğitimde öz-düzenleme öğretiminin önemi üzerine bir çalışma [A study on the importance of self-regulation teaching in education]. *Mehmet Akif Ersoy University Journal of Social Sciences Institute, 3*(5), 1-11.

Demir, R. Öztürk, N. & Dökme, İ. (2012). İlköğretim 7. sınıf öğrencilerinin fen ve teknoloji dersine yönelik motivasyonlarının bazı değişkenler açısından incelenmesi [Investigation of 7th Grade Primary School Students' Motivation towards Science and Technology Course in Terms of Some Variables]. *Mehmet Akif Ersoy University Journal of Social Sciences Institute, 12*(23), 1-21.

Demirbaş, M. & Yağbasan, R. (2004). Fen bilgisi öğretiminde, duyuşsal giriş özelliklerinin değerlendirilmesinin işlevi ve öğretim süreci içinde, öğretmen uygulamalarının analizi üzerine bir araştırma [A research on the progress of evaluating affective characteristics in science teaching and the analysis of teachers' practices in teaching process]. *Gazi Üniversitesi Kırşehir Eğitim Fakültesi Dergisi [Gazi University Journal of Kırşehir Education Faculty]*, *5*(2), 177-193.

Demirci, N. (2004). Öğrencilerin fiziğe giriş dersine karşı tutumları [Students' attitudes toward introductory physics Course]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [Hacettepe University Journal of Education], 26*, 33-40.

Eagly, A.H., & Chaiken, S. (1993). *The Psychology of Attitudes*. Fort Worth, TX: Harcourt Brace Jovanovich.

Gardner, P.L. (1975). Attitudes to science: A review. *Studies in Science Education, 2*, 1-41.

Güngör-Abak, A. & Eryılmaz, A. (2006). Gender differences in freshmen's physics related affective characteristics. *GIREP Conference 2006: Modeling in physics and physics education.* Amsterdam, Netherlands.

Hamurcu, H. (2002). Fen bilgisi öğretiminde etkili tutumlar [Effective attitudes in teaching science]. *Eğitim Araştırmaları Dergisi [Eurasian Journal of Educational Research], 8,* 144-152.

Hançer, H.A. (2008). Fen bilgisi öğretmen adaylarının fizik dersine yönelik tutumları [The attitudes of student teachers of science towards physics lesson]. *Çağdaş Eğitim Dergisi [Journal of Contemporary Education], 33*(354), 11-18.

House, J.D. & Prion, S.K. (1998). Student attitudes and academic background as predictors of achievement in college English. *International Journal of Instructional Media, 25*(1), 29-42.

Karasar, N. (2009). *Bilimsel Araştırma Yöntemi*. (21. Bas.). Ankara: Nobel Yayın Dağıtım.

Kauffman, D.F. (2004). Self-regulated learning in web-based environments: Instructional tools designed to facilitate cognitive strategy use, metacognitive processing and motivational beliefs. *Journal of Educational Computing Research, 30*,139-161.

Kaya, H. & Böyük, U. (2011). Attitude towards physics lessons and physical experiments of the high school students, *European Journal of Physics Education*, *2*(1), 38-49.

Kind, P., James, K., & Barmby, P. (2007). Developing attitudes towards science measures. *Internatial Journal of Science Education*, *29*(7), 871-893.

Koballa, R. JR. (1988). Attitude and related concepts in science education. *Science Education, 72*(2), 115-126.

Lowery, L.R., Bowyer, J., & Padilla, M.J. (1980). The Science Curriculum improvement study and student attitude. *Journal of Research in Science Technology*, *17*, 327-355.

MEB (2013). *Lise Fizik Dersi (9, 10, 11 ve 12. Sınıflar) Öğretim Programı*, Ankara.

Mujtaba, T. & Reiss, M.J. (2013) What sort of girl wants to study physics after the age of 16? Findings from a large-scale UK survey, *International Journal of Science Education, 35*(17), 2979-2998.

Murphy, P. & Whitelegg, E. (2006) Girls and physics: continuing barriers to 'belonging', *The Curriculum Journal*, *17*(3), 281-305.

Nuhoğlu, H. & Yalçın, N. (2004). The development of attitude scale for physics laboratory and the assessment of preserves teachers' attitudes towards physics laboratory. *Gazi University Journal of Kırşehir Education Faculty, 5*(2), 317-327.

Nuhoğlu, H. (2008). The development of an attitude scale for science and technology course. *Elementary Education Online, 7*(3), 627-639.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: a review of the literature and its implications. *International Journal of Science Education, 25*(9), 1047-1049.

Özyürek, A. & Eryılmaz, A. (2001). Factors affecting students towards physics. *Education & Science, 26*(120), 21-28.

Pell, T. & Jarvis, T. (2001). Developing attitude to science scales for use with children of ages from 5 to 11. *International Journal of Science Education*, *23*(8), 847-862.

Pendergast, E., Lieberman-Betz, R.G. & Vail, C.O. (2017). Attitudes and Beliefs of Prekindergarten Teachers Toward Teaching Science to Young Children, *Early Childhood Education Journal*, 45(1), 43-52.

Petty, R. (1995). *Attitude Change*. In A. Tesser (Ed.), Advanced Social Psychology. New York: NY, McGraw-Hill.

Pintrich, P.R. & De Groot, E.V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33-40.

Pintrich, R.R. (2000). *The Role of Goal Orientation in Self-Regulated Learning*. (Eds: M. Boekaerts, P.R.Pintrich, M.Zeidner), Handsbook of Self-Regulation. San Diego, CA: Academic Press.

Reid, N. & Skryabina, E.A. (2002) Attitudes towards physics, *Research in Science & Technological Education*, *20*(1), 67-81.

Reid, N. (2006). Thoughts on attitude measurement. *Research in Science & Technological Education*. *24*(1), 3-27.

Safran, M. (1993). Değişik öğretim basamaklarında tarih dersine ilişkin tutumlar üzerine bir araştırma. *Eğitim Dergisi*, *4*, 35-40.

Sarı, M. (2015). Teknik bilimler meslek yüksek okulu öğrencilerinin fizik dersine ilişkin düşünceleri ve fizik dersinden başarısızlıklarını olumsuz etkileyen faktörlerin öğrenci görüşlerine göre değerlendirilmesi [Vocational technical high school students of physical science basis of course views and physics course evaluation by failure from the negative factors affecting students' views]. *Eğitim ve Öğretim Araştırmaları Dergisi [Journal of Research in Education and Teaching]*, *4*(3), 206-210.

Schibeci, R.A. (1983). Selecting appropriate attitudinal objectives for school science. *Science Education*, *67*(5), 595-603.

Selvi, K. (1996). Tutumların ölçülmesi ve program değerlendirme [Measurement attitudes and curriculum evaluation]. *Anadolu Üniversitesi Eğitim Fakültesi Dergisi [Anadolu University Journal of Education Faculty]*, 6(2), 39-53.

Serin, O., Kesercioğlu, T., Saracaloğlu, A.S. & Serin, U. (2003). Sınıf öğretmenliği ve fen bilgisi öğrencilerinin fen (bilimleri)'e yönelik tutumları [The attitudes of the students in the primary school teaching and science programs towards science]. *Marmara Üniversitesi Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi [Marmara University Atatürk Education Faculty Journal Educational Sciences]*, *17*, 75–86.

Sorge, C. (2007). What happens? Relationship of age and gender with science attitudes from elementary to middle school. *Science Educator*, *16*(2), 33-37.

Stephens, K.R. (1999). Factors Affecting Science Related Attitudes in Academically Talented Youth. Unpublished Doctoral Dissertation. The University of Southern Mississippi.

Şengören, K.S., Tanel, R. & Kavcar, N. (2006). Optik dersine yönelik tutum ölçeği geliştirilmesi [The development of an attitude scale towards optics course]. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi [Pamukkale University Journal of Education]*, *20*, 63-68.

Tekbıyık, A. & Akdeniz, A.R. (2010). Ortaöğretim öğrencilerine yönelik güncel fizik tutum ölçeği: Geliştirilmesi, geçerlik ve güvenirliği [A Contemporary physics attitude scale for secondary school students: development, validity and reliability], *Türk Fen Eğitimi Dergisi [Journal of Turkish Science Education]*, *7*(4), 134-144.

Tekin, H. (1996). Eğitimde Ölçme ve Değerlendirme. Ankara: Yargı Yayınları.

Turgut, M.F. (1997). Eğitimde Ölçme ve Değerlendirme Metotları. Ankara.

Uzun, N. & Keleş, Ö. (2012). İlköğretim öğrencilerinin fen öğrenmeye yönelik motivasyon düzeylerinin değerlendirilmesi [Evaluation of primary school students' motivation levels for science learning. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi [Mustafa Kemal University Journal of Social Sciences Institute]*, *9*(20), 313-327.

Ülgen, G. (1997). Eğitim Psikolojisi, Kavramlar, İlkeler, Yöntemler, Kuramlar ve Uygulamalar. Ankara: Kurtiş Matbaası.

Üredi, I. (2005). *Algılanan anne baba tutumlarının ilköğretim 8. Sınıf öğrencilerinin öz-düzenleme stratejileri ve motivasyonel inançları üzerindeki etkisi [The contributions of*

perceived parenting styles to 8th class primary school students' self regulated learning strategies and motivational beliefs]. Yayınlanmamış Doktora tezi [[Unpublished doctorate dissertation]. Yıldız Teknik Üniversitesi [Yıldız Technical University], Sosyal Bilimler Enstitüsü [Social Sciences Institute], İstanbul.

Whitelegg, E. & Parry, M. (1999). Real-life contexts for learning physics: meanings, issues, and practice. *Physics Education, 34*, 68-72.

Yamaç, A. (2011). *İlköğretim Beşinci Sınıf Öğrencilerinin Öz-Düzenleyici Öğrenme Stratejileri İleMatematiğe Yönelik Tutum ve Başarıları Arasındaki İlişkilerin İncelenmesi [Examination of the relationships between primary fifth graders' self-regulated learning strategies and attitudes toward mathematics and mathematics achievement]*. Yayınlanmamış Yüksek lisans Tezi [Unpublished Master Thesis]. Afyon Kocatepe Üniversitesi [Afyon Kocatepe University], Sosyal Bilimler Enstitüsü [Social Sciences Institute], Afyon.

Yaman, S. & Dede, Y. (2007). Öğrencilerin fen ve teknoloji ve matematik dersineyönelik motivasyon düzeylerinin bazı değişkenler açısından incelenmesi [Examination of motivation level of students towards science and mathematics by some variables]. *Kuram ve Uygulamada Eğitim Yönetimi [Educational Administration: Theory and Practice], 52*, 615-638.

Yenice, N. Saydam, G. & Telli, S. (2012). İlköğretim öğrencilerinin fen öğrenmeye yönelik motivasyonlarını etkileyen faktörlerin belirlenmesi [Determining factors effecting on primary school students' motivation towards science learning]. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi [Ahi Evran University Journal of Kırşehir Education Faculty, 13*(2), 231-247.

Yeşilyurt, M. (2004). Student teachers' attitudes about basic physics laboratory, *Turkish Online Journal of Educational Technology, 3*(4), 49-57.

Zhang, P., Ding, L. & Mazur, E. (2017). Peer Instruction in introductory physics: A method to bring about positive changes in students' attitudes and beliefs. *Physical Review Physics Education Research, 13*(1), 010104.

Zacharia, Z. (2003). Beliefs, attitudes, and intentions of science teachers regarding the educational use of computer simulations and inquiry-based experiments in physics. *Journal of Research in Science Teaching, 40*(8), 792-823.

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Scaling of Ideal Teachers Characteristics with Pairwise Comparison Judgments According to Pre-service Teachers Opinions

Metin Yasar

To cite this article: Yasar, M. (2018). Scaling of Ideal Teachers Characteristics with Pairwise Comparison Judgments According to Pre-service Teachers Opinions. *International Journal of Assessment Tools in Education*, 5(1), 130-145. DOI: 10.21449/ijate.369233

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 130-145

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Scaling of Ideal Teachers Characteristics with Pairwise Comparison Judgments According to Pre-service Teachers Opinions

Metin Yasar

Educational Measurement and Evaluation, Pamukkale University, Denizli, Turkey

Abstract: In this study, scaling the characteristics that should be found in an ideal teacher according to the pre-service teachers by using the pairwise comparison method was aimed. Thirteen characteristics that an ideal teacher should have were given to 211 pre-service teachers in the working group, and these 13 properties were first asked to be considered as a whole, and then each property was asked to be compared to another property, one by one, to prefer one property to another. The research data were obtained from 211 pre-service teachers in fall semester of the 2015-2016 academic year. The data were scaled according to the pairwise comparison method. According to the findings obtained, when the characteristics were aligned from the most important characteristic that an ideal teacher should have according to the pre-service teachers to the most unimportant one, it was determined that; he/she should have an intellectual personality (U10) should have a sense of humor (U7), should be open to being criticized (U2), should be motivating (U1), should have a smiling expression (U5), should have a good usage of diction (U8), should be trustworthy (U3), should be creative (U6), should be a researcher (U9), should use teaching techniques well (U10), should give importance to the students (U4), should have good communication skills (U9), should keep the distance with the students (U12).

ARTICLE HISTORY

Received: 31 October 2017
Revised: 12 December 2017
Accepted: 14 December 2017

KEYWORDS

Scaling,
Pairwise Comparison,
Teacher Characteristics

1. INTRODUCTION

Nowadays it seems that the education system is in a student centered concept rather than a teacher centered one. Student centered education does not make a teacher insignificant, on the contrary it gives the teacher a more significant role. The most significant role of a teacher in education system is to assist the cognitive, affective and psychomotor development of students. An ideal teacher is a guide who takes care of all the students in class and enables required behavioral changes in the students by encouraging them to participate in class. The increase in the expectations of societies in education and by means of that in teachers switched the role of teachers in education system (Şahin, 2001), moreover the personality characteristics became more important. Along with the characteristics which a teacher is required to have such as being friendly, enthusiastic, in favor of change and progress, humanist, thinker and a person

Corresponding Author E-mail: myasar@pau.edu.tr

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.369233**

who expresses their own opinions (Brophy & Alleman, 1991), a teacher is also expected to be a person who communicates with students effectively (Bilen, 1995), a teacher who enables students to participate in the teaching – learning process effectively therefore helps them obtain behavioral change as a qualified teacher in the field, a teacher who utilizes convenient methods and instruments in order to meet educational needs (Şahin, 2001; Woolfolk, 1998), who listens to the problems of their students, who understands their students truly and tries to find solutions to their problems and a teacher who treats them as a friend (Ergün, Duman, Y. Kıncal & Arıbaş,1999).

When the studies which define the characteristics of a teacher are investigated, these characteristics come forward: Creativity, emotional adaptation, performing positive approaches towards students, positive attitudes towards teaching, socially good relationships, using the mother tongue efficiently, being sensitive, being able to develop empathy, avoiding judgements and participating in the social occasions of the society where they live (Confery, 1990; Good & Grouws, 1979; Rosenshine & Stevens, 1986; Ryan, 1960).

Since the target audience of teachers is students, they are required to have such characteristics as: enabling students to discover their potential and by developing their potential guiding them to their self-actualization, providing them the knowledge and the skills that would help them solve the real-world problems, establishing health relationships in order to prepare them for life, making the students trust, being gracious to them, being creative, caring about the students, being motivating, being open to criticism, being humoristic, having a good diction, having high communication skills, being sophisticated, utilizing teaching methods efficiently, being open and respectful to individual differences and being an enquirer. Unfortunately, claiming that all these characteristics are present in the teachers at a desired level is hardly possible. It is sometimes necessary to know the differences between the perceived and actual sizes of teacher qualifications mentioned above. The main purpose of the scale obtained from the difference or the correlation between perceived and actual size of the desired qualifications or any other variable is to put forward the methods of transition from empirical relationships based on observations to formal relationships based on rules (Anıl & Güler, 2006; Kan, 2008; Kart & Gelbal, 2014; Turgut & Baykul, 1992). Anıl and Güler (2006) perceived scaling in measuring process as a significant factor of the transition from the observations which shows qualitative distinction to the scales which show quantitative distinction. On the contrary, Stevens (1966) perceived scaling as marking objects with numbers based on a certain rule, testing hypothesis, determining whether a status or a concept is unidimensional or multidimensional and it was expressed that the most known reason of him to use scaling is grading (as cited in Anıl & Güler, 2006).

The approaches used in scaling are classified into two groups. The first of them is the approaches based on judge decisions and the second one is the approaches based on the reactions of test subjects. The classification of scaling approaches is given on Figure 1.

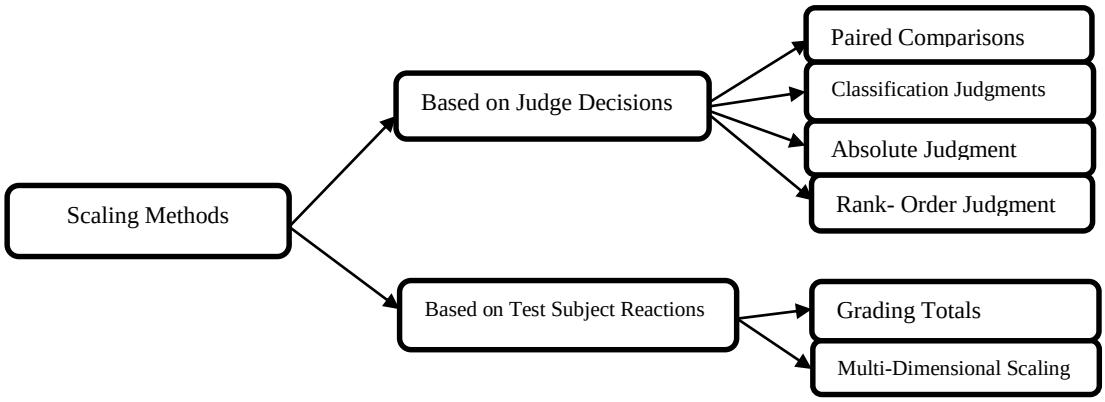


Figure 1. Main Approaches used in scaling (Arık & Kutlu, 2013).

The scaling approach based on judge decisions is to scale present stimulants at a determined level according to the judgments of observers and experts and in experimental methods, N number observers are demanded to determine stimulus levels of each of K number stimulants according to a certain method (Anıl & Güler, 2006; Turgut & Baykul, 1992; Yaşar, 2016). The size of the stimulants which are given to the observers is asked to be determined by comparing them to other stimulants. Therefore, the mean value of the judgments of observers gives the scale value of the stimulant.

In the approaches based on test subject reactions, it is not defined as stimulant centered but answerer centered approach. According to this approach, each answerer is placed somewhere on the scale according to the answers (reactions) that they give for the items (Crocker & Algina, 1986; as cited in Arık & Kutlu, 2013). Despite the rareness, it is obvious that the number of studies made on this subject is increasing. When the studies which were made considered, paired comparison method was used in order to scale the characteristics that a qualified teacher was required to have (Anıl & Güler, 2006), in order to scale the importance levels of professional teaching knowledge lessons (Nartgün, 2006), and to determine what characteristics the students who applied for a master's degree were required to have according to instructors (Güler & Anıl, 2009).

Attitude scale on addictive drugs was used in order to find out whether the scaling methods based on classification and sorting judgments gave similar results (Kan, 2008). The studies which were made also contained the scaling study on reliability and validity of field choice inventory of the senior students in the faculty of education (Öğretmen, 2008), overall impression, grading key, and the study of psychometric characteristics of three different evaluation methods based on the data collected from the compositions which were graded by Thurstone paired comparison method (Ömür, 2009), the scaling of the factors which were thought to be effective in placement test success with rank-order law (Bal, 2011).

Apart from these studies above, the studies which were also investigated are listed below: which characteristic competence of preservice teachers is more significant in the competence codes of teaching which were determined by Ministry of Education (Özer & Acar, 2011), the study to determine the consistency among scaling values obtained by scaling based on classification judgments and scaling based on test subject reactions (Öztürk, Özdemir & Gelbal, 2011), ranking judgment based scaling the characteristics which are thought to affect the academic success (Yaşar, 2016), ranking judgment based scaling of the mate selection criteria of university students (Bozgeyikli & Toprak, 2013), the investigation of the empathetic approach of elementary school administrators towards the professional problems of teachers with paired comparisons method (Ekinci, Bindak & Yıldırım, 2012), comparing the consistency

of the scale values obtained from scaling approaches based on paired comparisons judgments and ranking judgments (Albayrak & Gelbal, 2012), a paired comparison scaling study on the duties of education inspector in Turkey (Bülbül & Acar, 2012), scaling the characteristics which affect the success of elementary school students with completely ordered paired comparisons (Kara & Gelbal, 2013), judge decision based scaling of the assessment and evaluation competence of teachers (Arık & Kutlu, 2013), comparison of the evaluations which were made by grading key, overall impression and paired comparison methods (Ömür & Erkuş, 2013), comparison of two scaling methods: Paired Comparison and Ranking judgments (Acar Güvendir & Özer Özkan, 2013), the factors that affect the attitudes of students towards maths lesson according to teacher opinions (Arıcı, 2013), determining the scientific research self-efficacy perceptions of preservice teachers with paired comparison scaling method (Kart & Gelbal, 2014), determining the assessment and evaluation methods and instruments primarily used by elementary school teachers with paired comparison scaling method (Altun & Gelbal, 2014), determining the social activity choices of preservice teachers with paired comparison scaling method (Polat & Göksel, 2014), scaling the professional teaching knowledge lessons which senior students of faculty of education took with ranking judgment law (Yalçın & Avşar, 2014), the study in which it was detected whether the scale values of the purpose of internet use of preservice teachers obtained based on paired comparison and ranking method (Albayrak Sarı & Gelbal, 2015), the study to determine the measuring instruments (Gülşah Şahin, Boztunç, Öztürk & Taşdelen Teker, 2015). When research studies done abroad based on paired comparison method are considered, these studies listed below used paired comparison scaling method (as cited in Nartgün, 2006): the values of people on forests (Neuman, 1993), the value tendencies of Europeans (Francis et al., 2001), the perceptions of students on different nations (Zevinet al., 1998), the priority of social problems on natural resources (USDA Natural Resources Conservation Service, 1997), the perception of psychiatric patients on society's perspectives on mental illnesses (Freidle et al., 2003), the determination of the crispness levels of different brand crisps (Courcoux et al., 2005).

In this study, *the characteristics that an ideal teacher is required to have* were determined by using the scaling from the "*Law of Comparative Judgement IV Case Full Data Matrix".*

2. METHOD

Since in this study, the findings obtained from the study group do not generalize to the population, this study is not only a quantitative study but also a basic research study.

2.1. Study Group

This study consists of 211 preservice teachers who were getting education at the faculty of education of Pamukkale University, Denizli, Turkey in 2015-2016 academic year. The range of the preservice teachers according to certain variables is given on Table 1.

Table 1. Range of the preservice teachers of the study group according to certain variables.

Variable		f	%
Gender	Female	175	82.9
	Male	36	17.1
Department	1 Primary School Teaching	71	33.6
	2 Preschool Education	83	39.3
	3 Psychological Counseling and Guidance	57	27.0
Program Type	1 Daytime Education	121	57.3
	2 Evening Education	90	42.7
Grade Level	2nd Grade	99	46.9
	3rd Grade	82	38.9
	4th Grade	30	14.2

The preservice teachers of the study group is consisted of 175 (82.9%) female and 36 (17.1%) male students. 121 (57.3%) of them are daytime education students and 90(42.7%) of them are evening education students. 99 (46.9%) of them are second grade, 82 (38.9%) of them are third grade and 30 (14.2%) of them are fourth grade students. 71 (33.6%) of them are from the department of primary school teaching, 83 (39.3%) of them are from the department of preschool education and 57 (27.1%) of them are from the department of psychological counseling and guidance.

2.2. Data Collection Tool

In order to constitute a data collection tool, firstly the preservice teachers were asked to make a list of *"the characteristics that an ideal teacher is required to have"*. According to the answers of the preservice teachers, these characteristics were determined as: *(U1) should be motivating, (U2) should be open to criticism, (U3) should be reassuring, (U4) should care about students, (U5) should be cheerful, (U6) should be creative, (U7) should be humoristic, (U8) should have a decent diction, (U9) should have good communication skills, (U10) should be sophisticated, (U11) should utilize teaching methods efficiently, (U12) should be open and respective to differences, (U13) should be a researcher*. Statements on these characteristics were applied to 211 preservice teachers of the research group and the data which were used in the study were collected.

2.3. Data Analysis

Each preservice teacher who participated in the study was asked to prefer a characteristic to another one via paired comparison of *the characteristics that an ideal teacher is required to have*. Since there were 13 statements in the data collection tool, (13x(13-1))/2=78 paired comparisons were made in total. The frequency values of each characteristic were determined according to this process. Frequency matrix was constituted according to the frequency values. After the frequency matrix created, the values in each cell of the frequency matrix were divided into the number of the people and (P) values were obtained and therefore ratio matrix was created. Later on, the Unit Normal Deviance Matrix was created by obtaining (Z) values which were equaled to ratio matrix (P) values with the use of Microsoft Excel. The mean of columns in the unit normal deviance matrix was calculated and the scale values were achieved. The starting point of axis (zero point) was moved to the smallest scale value to determine the locations of the scale values on numerical axis (Anıl & Güler, 2006; Ekici, Bindak & Yıldırım, 2012; Turgut & Baykul, 1992).

2.4. Determination of the internal consistency of scale values

The internal consistency of scaling was examined in order to check whether the individuals of the group study behaved carefully on the statements of paired comparisons which they made for the stimulants. In order to determine the internal consistency of scale values, the concordance level of the observed p_{jk} rates with the p'_{jk} rates which are obtained from scale values (expected from the scaling) is considered (Turgut & Baykul, 1992). In order to examine the internal consistency, the concordance between theoretical ratios and observed ratios is investigated by creating a Z' unit normal deviation matrix and theoretical ratio matrix obtained from this matrix according to the scale values obtained from the data. In order to test the concordance level, formula (1.1) was used.

$$ME = \frac{\Sigma \left| P_{jk} - P'_{jk} \right|}{K\,(K-1)} \qquad (1.1)$$

ME: The mean value of the difference between theoretical ratios and observed ratios (mean error)

P_{jk}: The ratio obtained from observed frequencies
P'_{jk}: Theoretical ratio
K: The number of the stimulants

A small mean value obtained from the formula above indicates that the scale values obtained according to the paired comparisons that the observers made are reliable whereas a high error value indicates that the judgments of the observers are not reliable.

In order to determine the reliability which means the internal consistency of achieved scale values via the paired comparisons that 211 preservice teachers made in the study group of this study *"the characteristics that an ideal teacher is required to have",* these processes listed below were applied respectively.

1st Step: A theoretical Z' unit normal deviation matrix is created as it is showed in Table 2 by using scale values. In order to determine the elements of Z' matrix, $Z'_{jk} = S'_j - S'_k$ formula is used.

Table 2. Theoretical Unit Normal Deviation Matrix Z' ($Z_{jk}=S_j-S_k$)

		U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13
		0,228	0,202	0,556	0,942	0,359	0,58	0,06	0,422	0,964	0,000	0,935	1,048	0,917
U1	0,228	0,000												
U2	0,202	0,026	0,000											
U3	0,556	-0,328	-0,354	0,000										
U4	0,942	-0,714	-0,740	-0,386	0,000									
U5	0,359	-0,131	-0,157	0,197	0,583	0,000								
U6	0,580	-0,352	-0,378	-0,024	0,362	-0,220	0,000							
U7	0,060	0,168	0,142	0,496	0,882	0,299	0,520	0,000						
U8	0,422	-0,194	-0,220	0,134	0,52	-0,060	0,158	-0,362	0,000					
U9	0,964	-0,736	-0,762	-0,408	-0,022	-0,610	-0,384	-0,904	-0,542	0,000				
U10	0,000	0,228	0,202	0,556	0,942	0,359	0,580	0,060	0,422	0,964	0,000			
U11	0,935	-0,707	-0,733	-0,379	0,007	-0,580	-0,355	-0,875	-0,513	0,029	-0,940	0,000		
U12	1,048	-0,820	-0,846	-0,492	-0,106	-0,690	-0,468	-0,988	-0,626	-0,084	-1,050	-0,113	0,000	
U13	0,917	-0,689	-0,715	-0,361	0,025	-0,560	-0,337	-0,857	-0,495	0,047	-0,920	0,018	0,131	0,000

2nd Step: $\boldsymbol{P'}$ matrix is created by finding P'_{jk} rates equaled to Z'_{jk} values of Z' matrix from one unit normal distribution table. The matrix is given in Table 3.

Table 3. Theoretical Ratios Matrix ($\boldsymbol{P'}$)

		U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13
		0,228	0,202	0,556	0,942	0,359	0,58	0,06	0,422	0,964	0,000	0,935	1,048	0,917
U1	0,228	0,000												
U2	0,202	0,512	0,000											
U3	0,556	0,371	0,366	0,000										
U4	0,942	0,239	0,229	0,348	0,000									
U5	0,359	0,448	0,436	0,579	0,719	0,000								
U6	0,580	0,363	0,352	0,492	0,641	0,413	0,000							
U7	0,060	0,568	0,556	0,689	0,810	0,618	0,698	0,000						
U8	0,422	0,425	0,413	0,552	0,699	0,477	0,563	0,351	0,000					
U9	0,964	0,230	0,222	0,341	0,492	0,271	0,352	0,184	0,294	0,000				
U10	0,000	0,591	0,579	0,712	0,826	0,641	0,719	0,523	0,662	0,832	0,000			
U11	0,935	0,239	0,233	0,352	0,501	0,281	0,359	0,189	0,305	0,512	0,174	0,000		
U12	1,048	0,206	0,198	0,312	0,457	0,244	0,319	0,161	0,264	0,468	0,147	0,456	0,000	
U13	0,917	0,245	0,236	0,359	0,512	0,288	0,367	0,195	0,309	0,519	0,179	0,508	0,551	0,000

Error matrix $p\left(p_{jk}-p'_{jk}\right)$ is created by the absolute value of the differences between observed ratios and theoretical ratios. The Error matrix is given in Table 4.

Table 4. Error Matrix

		U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13
		0,228	0,202	0,556	0,942	0,359	0,58	0,06	0,422	0,964	0,000	0,935	1,048	0,917
U1	0,228	0,000												
U2	0,202	0,005	0,000											
U3	0,556	0,085	0,018	0,000										
U4	0,942	0,084	0,089	0,063	0,000									
U5	0,359	0,056	0,004	0,038	0,023	0,000								
U6	0,580	0,022	0,033	0,021	0,093	0,021	0,000							
U7	0,060	0,016	0,004	0,029	0,051	0,007	0,006	0,000						
U8	0,422	0,015	0,025	0,062	0,086	0,056	0,033	0,04	0,000					
U9	0,964	0,023	0,028	0,041	0,077	0,014	0,091	0,017	0,030	0,000				
U10	0,000	0,070	0,048	0,074	0,083	0,017	0,073	0,04	0,014	0,038	0,000			
U11	0,935	0,030	0,069	0,019	0,045	0,035	0,035	0,080	0,061	0,083	0,047	0,000		
U12	1,048	0,016	0,034	0,009	0,030	0,005	0,026	0,009	0,065	0,152	0,000	0,007	0,000	
U13	0,917	0,047	0,052	0,045	0,004	0,051	0,298	0,056	0,053	0,131	0,042	0,039	0,017	0,000
	Total	0,469	0,404	0,401	0,492	0,206	0,562	0,242	0,223	0,404	0,089	0,046	0,017	0,000

Mean error is found by finding the total of the column totals of error matrix given in Table 4 and dividing it into K.(K-1) number. For this study the mean error ratio was calculated as:

$$ME = \frac{\Sigma\left|P_{jk} - P'_{jk}\right|}{K\,(K-1)} = \frac{3.555}{13\,(13-1)} = 0{,}022$$

This value may be accepted as a considerably small value. The case that mean error ratio value is considerably small shows that scale values have internal consistency.

3. FINDINGS

In this part of the study, paired comparisons and interpretations of *the characteristics that an ideal teacher is required to have* were given according to the gender, program type and grades of preservice teachers. Here, how many times the characteristic in the line was chosen compared to the character in the column; i. line and j. column element (U_{ij}), by the preservice teachers. According to this, it was seen that) =104 for U1 U2 characteristics. This means that the number of preservice teachers who preferred U1 to U2 is 104 out of 211. Likewise, the number of preservice teachers who preferred U2 characteristic to U1 is [(U2, U1) = n - (U1, U2)] = 211-104 = 107.

Table 5. The Raw Scores Matrix of the Preservice teachers [F]

	STIMULANTS (U_j)												
U_i	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13
U1		104	153	182	129	132	87	126	171	70	158	168	153
U2	107		133	185	119	132	92	120	174	77	151	166	154
U3	58	78		153	96	103	68	106	150	42	135	61	128
U4	29	26	58		51	93	46	78	124	50	114	121	102
U5	82	92	115	160		121	78	123	160	77	147	164	143
U6	79	79	108	118	90		62	98	159	71	144	152	138
U7	124	119	143	165	133	149		146	173	108	160	179	162
U8	85	91	105	133	88	113	65		146	71	163	172	160
U9	40	37	61	87	51	52	38	65		39	120	145	129
U10	141	134	169	161	134	140	103	140	172		168	184	168
U11	53	60	76	97	64	67	51	48	91	43		114	112
U12	43	45	61	90	47	59	32	39	66	27	97		90
U13	58	57	83	109	68	138	49	51	82	43	99	121	
total	899	922	1265	1640	1070	1299	771	1140	1668	718	1656	1747	1639

Ratio (P) matrix was created by dividing the values of judgements located in the each cell of Frequency (F) matrix into the number of the judges (N=211). The ratios (P) matrix is given in Table 6. Since the ratio values of ratio matrix are symmetrical to main diagonal, the sum of the ratios is equal to 1.

Table 6. Ratio Matrix (P)

	STIMULANTS (U_j)												
U_i	**U1**	**U2**	**U3**	**U4**	**U5**	**U6**	**U7**	**U8**	**U9**	**U10**	**U11**	**U12**	**U13**
U1		0,492	0,725	0,862	0,611	0,625	0,412	0,597	0,810	0,331	0,748	0,796	0,725
U2	0,507		0,630	0,876	0,563	0,625	0,436	0,568	0,824	0,364	0,715	0,786	0,729
U3	0,274	0,369		0,725	0,454	0,488	0,322	0,502	0,710	0,199	0,639	0,289	0,606
U4	0,137	0,123	0,274		0,241	0,440	0,218	0,369	0,587	0,236	0,540	0,573	0,483
U5	0,388	0,436	0,545	0,758		0,573	0,369	0,582	0,758	0,364	0,696	0,777	0,677
U6	0,374	0,374	0,511	0,559	0,426		0,293	0,464	0,753	0,336	0,682	0,720	0,654
U7	0,587	0,563	0,677	0,781	0,630	0,706		0,691	0,819	0,511	0,758	0,848	0,767
U8	0,402	0,431	0,497	0,630	0,417	0,535	0,308		0,691	0,336	0,772	0,815	0,758
U9	0,189	0,175	0,289	0,412	0,241	0,246	0,180	0,308		0,184	0,568	0,687	0,611
U10	0,668	0,635	0,800	0,763	0,635	0,663	0,488	0,663	0,815		0,796	0,872	0,796
U11	0,251	0,284	0,360	0,459	0,303	0,317	0,241	0,227	0,431	0,203		0,540	0,530
U12	0,203	0,213	0,289	0,426	0,222	0,279	0,151	0,184	0,312	0,127	0,459		0,426
U13	0,274	0,270	0,393	0,516	0,322	0,654	0,232	0,241	0,388	0,203	0,469	0,573	
total	4,261	4,370	5,995	7,773	5,071	6,156	3,654	5,403	7,905	3,403	7,848	8,280	7,768

(Z) standard values equaled to the cell values (P) of ratios matrix was found and unit normal deviation matrix in Table 7 was obtained. In the unit normal deviation matrix (Z), the elements are opposite signed according to main diagonal but their values are absolute. The column values of the stimulants in the unit normal deviation matrix (Z) were summed up. The column sums in the matrix were divided into the numbers of elements in the column and the scale values of the stimulants were calculated. The scale values are given in Table.7.

Table 7. Unit Normal Deviation Matrix (Z Matrix)

	STIMULANTS (Uj)												
U_i	**U1**	**U2**	**U3**	**U4**	**U5**	**U6**	**U7**	**U8**	**U9**	**U10**	**U11**	**U12**	**U13**
U1		-0,018	0,598	1,092	0,283	0,320	-0,222	0,246	0,879	-0,435	0,671	0,828	0,598
U2	0,018		0,333	1,159	0,161	0,320	-0,161	0,173	0,933	-0,345	0,570	0,795	0,612
U3	-0,598	-0,333		0,598	-0,113	-0,030	-0,461	0,006	0,556	-0,845	0,358	-0,556	0,271
U4	-1,092	-1,159	-0,598		-0,701	-0,149	-0,779	-0,333	0,222	-0,716	0,101	0,185	-0,042
U5	-0,283	-0,161	0,113	0,701		0,185	-0,333	0,209	0,701	-0,345	0,515	0,763	0,461
U6	-0,320	-0,320	0,030	0,149	-0,185		-0,542	-0,089	0,686	-0,422	0,475	0,584	0,396
U7	0,222	0,161	0,461	0,779	0,333	0,542		0,501	0,915	0,030	0,701	1,029	0,732
U8	-0,246	-0,173	-0,006	0,333	-0,209	0,089	-0,501		0,501	-0,422	0,747	0,897	0,701
U9	-0,879	-0,933	-0,556	-0,222	-0,701	-0,686	-0,915	-0,501		-0,897	0,173	0,488	0,283
U10	0,435	0,345	0,845	0,716	0,345	0,422	-0,030	0,422	0,897		0,828	1,136	0,828
U11	-0,671	-0,570	-0,358	-0,101	-0,515	-0,475	-0,701	-0,747	-0,173	-0,828		0,101	0,077
U12	-0,828	-0,795	-0,556	-0,185	-0,763	-0,584	-1,029	-0,897	-0,488	-1,136	-0,101		-0,185
U13	-0,598	-0,612	-0,271	0,042	-0,461	0,396	-0,732	-0,701	-0,283	-0,828	-0,077	0,185	
$\sum Z$	-4,225	-4,568	0,035	5,060	-2,527	0,352	-6,406	-1,710	5,346	-7,191	4,960	6,436	4,733
$\bar{Z}$	-0,325	-0,351	0,003	0,389	-0,194	0,027	-0,493	-0,132	0,411	-0,553	0,382	0,495	0,364
S_C	0,228	0,202	0,556	0,942	0,359	0,580	0,060	0,422	0,964	0,000	0,935	1,048	0,917

13 characteristics that an ideal teacher is required to have according to preservice teachers, the scale values obtained by the law of paired comparisons and the stimulant rank values of the characteristics are displayed in Table 8.

The significance order of the characteristics that an ideal teacher is required to have were determined considering the gender, the program type (daytime education- evening education)

and the grades of the preservice teachers. The findings obtained according to these characteristics are displayed in Table 8.

When the findings in Table 8 are considered in general, the most important characteristic was stated as *(U10) should be sophisticated*, and the others were ranked respectively as *(U7) should be humoristic, (U2) should be open to criticism, (U1) should be motivating, (U5) should be cheerful, (U8) should have a decent dictation, (U3) should be reassuring, (U6) should be creative, (U13) should be a researcher, (U11) should utilize teaching methods efficiently, (U4) should care about students, (U9) should have good communication skills, (U12) should keep distance from students.*

On the other hand, when the gender of the preservice teachers is considered, *the characteristics that an ideal teacher is required to have* are ordered as: the most significant characteristic according to both male and female preservice teachers is *a teacher should be sophisticated*, the second significant characteristic according to male preservice teachers is *a teacher should be humoristic* whereas this characteristic is the third significant according to female preservice teachers. While the most significant characteristic according to male preservice teachers is *a teacher should be open to criticism*, according to female preservice teachers this characteristic is the second significant characteristic. The fourth significant characteristic according to both male and female preservice teachers is *a teacher should be motivating.* Similarly, according to both male and female preservice teachers the least significant characteristic is *a teacher should keep distance from students.*

When the school type (daytime and evening education) is considered, the most significant characteristic that an ideal teacher is required to have is *a teacher should be humoristic* according to the preservice teachers of daytime education, whereas according to the evening education preservice teachers this characteristic is the second significant one. The most significant characteristic according to the evening education preservice teachers is *a teacher should be sophisticated*, however, this characteristic is the second significant characteristic according to the daytime education preservice teachers. Furthermore, the characteristic of *a teacher should keep distance from students* is the least significant one according to both daytime and evening education preservice teachers.

When the grades of the preservice teachers are considered, the paired comparison results based on the significance ranks of the characteristics that an ideal teacher is required to have are stated as: the characteristic of *a teacher should be sophisticated* is the most significant characteristic according to the second and third grade preservice teachers, but according to the fourth grade preservice teachers this characteristic is the fourth significant one. On the other hand, the most significant characteristic according to the fourth grade preservice teachers is *a teacher should be open to criticism*, while this characteristic is the third significant characteristic according to the second graders and the fourth significant characteristic according to the third graders. According to the preservice teachers of all grades, the characteristic *a teacher should keep distance from students* is the least significant one.

(U4) a teacher should care about students characteristic, which is indeed supposed to be among the most significant characteristics, is the eleventh according to the second grade preservice teachers and the tenth according to the third and the fourth grade preservice teachers. Similarly, *(U11) a teacher should utilize the teaching methods efficiently* characteristic which can be seen as a significant characteristic in the professional development of a teacher, is the tenth according to the second grade preservice teachers, the eleventh according to the third grade preservice teachers and the eighth according to the fourth grade preservice teachers. Likewise, all three grades of preservice teachers stated that the least significant characteristic in the scale is *(U12) a teacher should keep distance from students.*

Table 8. The scale values and stimulant ranks of "the characteristics that an ideal teacher is required to have" according to the general, gender, school type and grades of preservice teachers.

The characteristics that an ideal teacher is required to have according to preservice teachers		Preservice Teachers (General)		Gender				School Type				Grade					
				Male Preservice Teachers		Female Preservice Teachers		Daytime Education		Evening Education		2nd Grade		3rd Grade		4th Grade	
		Scale Values	Stimulant Ranks	**Scale Values**	Stimulant Ranks	**Scale Values**	Stimulant Ranks	**Scale Values**	Stimulant Ranks	**Scale Values**	Stimulant Ranks	**Scale Values**	Stimulant Ranks	**Scale Values**	Stimulant Ranks	**Scale Values**	Stimulant Ranks
U1	Should be motivating	0,228	**4**	0,247	**4**	0,164	**3**	0,102	**3**	0,337	**4**	0,325	**4**	0,155	**3**	0,026	**2**
U2	Should be open to criticism	0,202	**3**	0,180	**2**	0,203	**4**	0,217	**4**	0,209	**3**	0,259	**3**	0,324	**4**	0,000	**1**
U3	Should be reassuring	0,556	**7**	0,702	**9**	0,513	**8**	0,447	**7**	0,756	**8**	0,662	**8**	0,569	**7**	0,456	**9**
U4	Should care about students	0,942	**11**	0,915	**12**	0,949	**9**	0,845	**10**	1,114	**11**	1,041	**11**	1,055	**10**	0,624	**10**
U5	Should be cheerful	0,359	**5**	0,482	**7**	0,330	**5**	0,247	**5**	0,573	**6**	0,489	**5**	0,340	**5**	0,238	**6**
U6	Should be creative	0,580	**8**	0,481	**6**	0,382	**6**	0,488	**8**	0,591	**7**	0,612	**7**	0,591	**8**	0,293	**7**
U7	Should be humoristic	0,060	**2**	0,202	**3**	0,027	**2**	0,000	**1**	0,188	**2**	0,146	**2**	0,034	**2**	0,103	**3**
U8	Should have a decent dictation	0,422	**6**	0,473	**5**	0,411	**7**	0,365	**6**	0,520	**5**	0,502	**6**	0,496	**6**	0,194	**5**
U9	Should have good communication skills	0,964	**12**	0,903	**11**	0,981	**11**	0,935	**12**	1,047	**10**	1,054	**12**	1,017	**9**	0,801	**12**
U10	Should be sophisticated	0,000	**1**	0,000	**1**	0,000	**1**	0,036	**2**	0,000	**1**	0,000	**1**	0,000	**1**	0,165	**4**
U11	Should utilize the teaching methods efficiently	0,935	**10**	0,684	**8**	0,989	**12**	0,851	**11**	1,090	**9**	0,990	**10**	1,204	**11**	0,397	**8**
U12	Should keep distance from students	1,048	**13**	1,014	**13**	1,156	**13**	1,041	**13**	1,292	**13**	1,177	**13**	1,323	**13**	0,846	**13**
U13	Should be a researcher	0,917	**9**	0,732	**10**	0,950	**10**	0,782	**9**	1,138	**12**	0,801	**9**	1,283	**12**	0,715	**11**

The characteristics that an ideal teacher is required to have according to preservice teachers were scaled by using paired comparisons method according to the gender, school type and grades of preservice teachers. Spearman's rho correlation method was utilized in order to determine whether there was a meaningful correlation between the results of paired comparisons which were made according to the mentioned characteristics of the preservice teachers. The results obtained are displayed in Table 9.

Table 9. Spearman's rho correlation coefficients of the significance levels of the characteristics that an ideal teacher is required to have according to the gender, program type (daytime education and evening education) and grades of the preservice teachers

			MPT	FPT	DE	EE	SG	TG	FG	GPT
Spearman' rho	MPT	Correlation Coefficient	1,000							
		Sig. (2-tailed)	.							
		N	13							
	FPT	Correlation Coefficient	,890**	1,000						
		Sig. (2-tailed)	,000	.						
		N	13	13						
	DE	Correlation Coefficient	,896**	,967**	1,000					
		Sig. (2-tailed)	,000	,000	.					
		N	13	13	13					
	EE	Correlation Coefficient	,967**	,929**	,929**	1,000				
		Sig. (2-tailed)	,000	,000	,000	.				
		N	13	13	13	13				
	SG	Correlation Coefficient	,956**	,962**	,978**	,956**	1,000			
		Sig. (2-tailed)	,000	,000	,000	,000	.			
		N	13	13	13	13	13			
	TG	Correlation Coefficient	,890**	,956**	,945**	,967**	,934**	1,000		
		Sig. (2-tailed)	,000	,000	,000	,000	,000	.		
		N	13	13	13	13	13	13		
	FG	Correlation Coefficient	,940**	,874**	,896**	,929**	,918**	,874**	1,000	
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	.	
		N	13	13	13	13	13	13	13	
	GPT	Correlation Coefficient	,940**	,951**	,984**	,951**	,995**	,940**	,907**	1,000
		Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	.
		N	13	13	13	13	13	13	13	13

** P<0,01

MPT: Male Preservice Teacher; FPT: Female Preservice Teacher; DE: Daytime Education; EE: Evening Education; SG: Second Grade; TG: Third Grade; FG: Fourth Grade; GPT: General Preservice Teachers

When Table 9 is investigated, the minimal value of correlation coefficients for N*(N-1)/2 paired comparisons made according to the gender, program type (daytime and evening education) and grades of the preservice teachers who participated in the study group is between FG and FPT (r = 0.874) and between FG and TG (r = 0.874) while the maximum correlation coefficient is between GPT and SG (r = 0.995).

In addition, the Spearman's rho correlation coefficients of the paired comparisons in Table 9 indicate a positively high level correlation and also it is clear that the correlation coefficients of paired comparison results are statistically significant at the 0.01 level.

4. DISCUSSION, RESULTS AND SUGGESTIONS

In this study, the perceptions of preservice teachers who were students at faculty of education on the characteristics that an ideal teacher is required to have and the characteristics which were assumed to be related were scaled via full data matrix with the use of the law of paired comparatives V case. The study was carried out on the data collected from 211 preservice teachers who were 2nd, 3rd and 4th grade students at Pamukkale University, Denizli, Turkey, faculty of education, department of primary school teaching (n=71; 33.6%),

department of preschool education (n=83; 39.3%), department of psychological counseling and guidance (n=57; 27.0%). The preservice teachers were asked to prefer one characteristic to another by making paired comparisons of 13 characteristics that an ideal teacher is required to have. After making paired comparisons, the frequency of each characteristic was determined. Frequency matrix was created with these frequencies. Then, the value of each cell in the frequency matrix was divided into the number of the participants in the study group (n=211) therefore ratios (***P***) matrix was created. Unit deviation matrix (Z) equaled to each (P) value of the ratios matrix was also created. In order to determine whether the paired comparison judgments of preservice teachers that they made for the stimulants given, internal consistency of scaling was examined. For this, the concordance level of observed $\boldsymbol{P_{jk}}$ ratios with $\boldsymbol{P'_{jk}}$ values obtained from the scale values (expected from the scale values) is examined (Turgut & Baykul, 1992). In order to examine the internal consistency, the concordance between observed ratios and theoretical ratios is checked by obtaining a Z' unit normal deviation matrix created from the scale values which were obtained by the data and a theoretical matrix out of this matrix.

Calculated mean error value can be accepted as a quite small value. A considerably small mean error value (ME=0.022<0.05) indicates that the scale values have internal consistency. The first question of the study was how the characteristics that an ideal teacher is required to have according to preservice teachers were ranked from the most significant characteristic to the least significant one. Therefore, the preservice teachers were asked to compare each characteristic to the others as pairs using the law of paired comparisons. According to the findings obtained, among 13 characteristics that an ideal teacher is required to have, *(U10) a teacher should be sophisticated* was stated as the most significant characteristic. On the other hand, the least significant characteristic was stated as *(U12) a teacher should keep distance from students*.

The second question of the study is whether the significance rank of the characteristics that an ideal teacher is required to have vary or not considering the gender of preservice teachers. The finding that was reached when the gender of preservice teachers was considered stated that the most significant characteristic for both male and female teachers was *(U10) a teacher should be sophisticated.* For male preservice teachers the characteristic of *(U2) a teache*r *should be open to criticism* was the second significant characteristic while the characteristic of *(U7) a teacher should be humoristic* was the second significant characteristic for female preservice teachers. According to both male and female preservice teachers, the characteristic of *(U12) a teacher should keep distance from students* was stated as the least significant characteristic among 13 characteristics that an ideal teacher is required to have.

The third question of the study is whether the significance rank of the characteristics that an ideal teacher is required to have vary or not according to the program type of preservice teachers. According to the preservice teachers of daytime education, the most significant characteristic that an ideal teacher is required to have was *(U7) a teacher should be humoristic*, whereas this characteristic was the second according to the preservice teachers of evening education. The most significant characteristic according to the preservice teachers of evening education was *(U10) a teacher should be sophisticated*, while this characteristic was stated as the second according to the preservice teachers of daytime education.

The fourth question of the study is whether the significance rank of the characteristics that an ideal teacher is required to have vary or not according to the grade of preservice teachers. When the grade of preservice teachers was considered, according to the second and the third grade preservice teachers, the most significant characteristic was *(U10) a teacher should be sophisticated*, while according to the fourth grade preservice teachers this characteristic was stated as the fourth significant characteristic. On the other hand, according to the fourth grade preservice teachers, the most significant characteristic was stated as *(U2) a*

teacher should be open to criticism, while this characteristic was stated as the third significant characteristic according to the second grade preservice teachers and the fourth significant characteristic according to the third grade preservice teachers. According to all grades of preservice teachers the least significant characteristic was determined as *(U12) a teacher should be open and respective to differences.*

A paired comparison scaling study on teacher qualifications was carried out by Anıl and Güler (2006). Apart from this research, no other study handling teacher qualifications has been observed. Anıl and Güler (2006) examined eight qualifications as teacher characteristics in their study. In their study, the most significant qualification was stated as *working with passion,* it was followed respectively by the qualifications as *having the skill of imparting knowledge, having good communication skills, being open to technological developments, having the content knowledge, being democratic, being open to criticism* and the least significant qualification was stated as *being humoristic.*

It is clear that the characteristics of this study are stated as the same with the *being humoristic, communication skills, being open to criticism* qualifications of the study of Anıl and Güler (2006) and yet the other variables are stated as different.

While *communication skills* was determined as the third most significant characteristics in the study of Anıl and Güler (2006), in this study it was ranked as the twelfth. In Anıl and Güler's (2006) study, the characteristic of *being open to criticism* was ranked as the seventh in terms of significance whereas in this study, this characteristic was ranked as the third. In the study of Anıl and Güler (2006), the characteristic of *being humoristic* was ranked as the last in terms of significance among eight characteristics, while in this study it was ranked as the second among thirteen characteristics. It is clear that in the study of Anıl and Güler (2006), the mutual characteristics are not in the same significance order.

In the study of Anıl and Güler (2006) there were eight qualifications of teachers within the research but in this study there were 13 characteristics. While Anıl and Güler (2006) studied by considering the judgments of university students in general terms, in this study apart from the general judgments of preservice teachers who participated in the study, the variables of their gender, grades and program types (daytime education and evening education) were taken into consideration.

It is easy to see the known fact that there are few studies about scaling when the literature review is done. Therefore, the need for more studies on this field emerges spontaneously. It is thought that the researchers who are willing to study on this field can work on the subjects such as the effectiveness of teaching.

5. REFERENCES

Acar Güvendir, M. & Özer Özkan, Y. (2013). İki ölçekleme yönteminin karşılaştırılması: ikili karşılaştırma ve sıralama yargıları [A comparison of two scaling methods: Pair wise comparison and rank-order judgments scaling]. *Journal of Educational Sciences Research*, 3 (1), 105–119.

Albayrak, A. & Gelbal, S. (2015). İkili karşılaştırmalar yargılarına ve sıralama yargılarına dayalı ölçekleme yaklaşımlarının karşılaştırılması [A Comparison of scaling procedures based on pair-wise comparison and rank-order judgments scaling]. *Journal of Measurement and Evaluation in Education and Psychology*, 6 (1), 126-141.

Arıcı, Ö. (2013). Öğretmen görüşlerine göre öğrencilerin matematik dersine yönelik tutumlarını etkileyen faktörlerin ölçeklenmesi çalışması [A Scaling study for the factors affect the attitudes of students towards maths lesson according to the views of teachers]. *Journal of Education Ege*, 14 (2), 25-40.

Anıl, D. & Güler, N. (2006). İkili karşılaştırma yöntemi ile ölçekleme çalışmasına bir örnek [an example of the scaling study by pair-wise comparison method]. *H.U. Journal of Education*, 30 (36), 30-36.

Altun, A. & Gelbal, S. (2014). Öğretmenlerinin kullandıkları ölçme ve değerlendirme yöntem veya araçlarının ikili karşılaştırma yöntemiyle belirlenmesi [Determining teachers' measurement tools or techniques via pair-wise comparison method]. *Journal of Measurement and Evaluation in Education and Psychology*. 5 (1), 1-11.

Arık, R. S. & Kutlu, Ö. (2013). Öğretmenlerin ölçme ve değerlendirme alanı yeterliklerinin yargıcı kararlarına dayalı ölçeklenmesi [Scaling primary school teachers' competence based on judgmental decisions in the field of measurement and evaluation]. *Journal of Educational Sciences Research*, 3 (2), 163-196.

Bilen, M. (1995). *Planlamadan Uygulamaya Öğretim* [Teaching from Planning to implementation], Ankara.

Bal, Ö. (2011). Seviye belirleme sınavı (SBS) başarısında etkili olduğu düşünülen faktörlerin sıralama yargıları kanunuyla ölçeklenmesi [The Scaling of the factors which are considered to be effective on the success in the level determination exam (LDE) by ranking judgement law]. *Journal of Measurement and Evaluation in Education and Psychology*, 2 (2), 200-209.

Bozgeyikli, H. & Toprak, E. (2013) Üniversiteli gençlerin eş seçimi kriterlerinin sıralama yargılarıyla ölçeklenmesi [University youth's mate selection criteria by Rank Order Judgement scaling]. *Journal of Youth Research*, 1(1), 68-87.

Bülbül, T. & Acar, M. (2012). A pair-wise scaling study on the missions of education supervisors in Turkey. *International Journal of Human Sciences, 9* (2), 623-640.

Brophy, J. E. & Alleman, J. (1991). Activities as instructional tools: A framework for analysis and evaluation. *Educational Researcher*, *20*, 9-23.

Confery, J. (1990). What constructivism ımplies for teaching. *In Constructivist views on the teaching and learning of mathematics*, ed. R.Davis, C.Maher,& N.Noddingo. Reston, VA: *National Council of Teachers of Mathematics.*

Ekinci, A., Bindak , R. & Yıldırım, C. (2012). İlköğretim okulu yöneticilerinin öğretmenlerin mesleki sorunlarına empatik yaklaşımlarının ikili karşılaştırmalar metodu ile incelenmesi [A research regarding the empathic approaches of school managers about professional problems of teachers by pair-wise comparisons method]. *Gaziantep University Journal of Social Sciences*, 11 (3), 759-776.

Ergün, M., Duman,T., Y.Kıncal, R., & S. Arıbaş (1999). İdeal bir öğretim elemanın özellikleri [Characteristics of an ideal instructor]. *Afyon Kocatepe University, Journal of Social Sciences*, 3, 1-11.

Good, T.L. & Grouws, D.A. (1979). The Missouri Mathematics Effectiveness Project. *Journal of Educational Pschology*, 71(79) 357-375.

Güler, N. & Anıl, D. (2009). Scaling through pair-wise comparison method in required characteristics of students applying for post graduate programs. *International Journal of Human Sciences*, 6 (1), 627-639.

Gülşah Şahin, M., Boztunç Öztürk, N. & Taşdelen Teker, G. (2015). Öğretmen adaylarının başarılarının değerlendirilmesinde tercih ettikleri ölçme araçlarının belirlenmesi [determining the pre-service teachers' measurement tool preferences for evaluation of their achievement]. *Journal of Measurement and Evaluation in Education and Psychology*, 6 (1), 95-106.

Heldsinger, S. & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37(2).

Kan, A.(2008). Yargıcı kararlarına dayalı ölçekleme yöntemlerinin karşılaştırılması üzerine amprik bir çalışma [An comprasion of scaling methods based on judge decisions: an empricial study]. *Hacettepe University Journal of Education*, 35, 186-194.

Kara, Y. & Gelbal, S. (2013). İlköğretim öğrencilerinin başarılarını etkileyen özelliklerin tam sıralama halinde ikili karşılaştırmalar yöntemiyle ölçeklenmesi [Scaling of the characteristics affecting the success of primary school students by the method of pairwise comparisons in totally rank order]. *Journal of Measurement and Evaluation in Education and Psychology*, 4(1), 33-51.

Kart, A., & Gelbal, S., (2014). Öğretmen adaylarının bilimsel araştırma öz yeterlik algılarının karşılaştırmalı yargılar yöntemiyle belirlenmesi. [Determining prospective teachers' self-efficacy perception on scientific skills via pair-wise comparison method]. *Journal of Measurement and Evaluation in Education and Psychology*, 5 (1), 12-23.

Nartgün, Z, (2006). Öğretmenlik meslek bilgisi derslerinin önem düzeyinin ikili karşılaştırmalarla ölçeklenmesi [Scaling of the importance level of professional teaching knowledge courses using pairwise comparisons]. *A.İ.B.Ü. Journal of Faculty of Education*, 6 (2), 161- 176.

Öğretmen, T. (2008). Alan tercih envanteri: Ölçeklenmesi, geçerliği ve güvenirliği [Subject preference ınventory: Scaling, validity and reliability]. *Journal of Turkish Educational Sciences*, 6 (3) 507-522.

Ömür, S. & Erkuş, A. (2013). Dereceli puanlama anahtarıyla, genel izlenimle ve ikili karşılaştırmalar yöntemiyle yapılan değerlendirmelerin karşılaştırılması [Comparison of the evaluations which were done with rubric, overall ımpression and paired comparisons]. *Hacettepe University Journal of Education*, 28 (2), 308-320.

Özer, Y. & Acar, M. (2011). (4) Öğretmenlik mesleği genel yeterlikleri üzerine ikili karşılaştırma yöntemiyle bir ölçekleme çalışması [A Scaling study using the method of pairwise comparisons on the general qualifications of the teaching profession]. *Cukurova University faculty of Education Journal*, 3 (40), 89-101.

Öztürk, N., Özdemir, S. & Gelbal, S. (2011). İki farklı ölçekleme yaklaşımından elde edilen ölçek değerleri tutarlılığının incelenmesi [Examining the reliability of scale values obtained from two different scaling approaches]. *20th National Educational Sciences Congress*, 8-10 September 2011. Burdur.

Polat, B. & Göksel, H.Ç. (2014). Öğretmen adaylarının sosyal aktivite tercihlerinin ikili karşılaştırmalı ölçekleme yöntemiyle belirlenmesi [Determination of candidate teachers' social activity preferences by pair-wise comparision scaling method]. *Journal of Measurement and Evaluation in Education and Psychology*, 5 (1), 88-100.

Rosenshine, B. & Stevens, R. (1986). Teaching Functions, *In Handbook of Research on Teaching*, ed. M.C. Wittrock, 3rd ed. New York: Macmillan

Ryan, D. (1960). *Characteristics of Effective Teachers.* Washington, DC: American Council on Education.

Şahin, A., (2001) Öğretmen algılarına göre etkili öğretmen davranışları [Effective teacher's attitudes according to teacher's perceptions]. *Ahi Evran University, Journal of Kırşehir education Faculty*, 12 (1), 239-259.

Turgut, M. F. & Baykul, Y. (1993). *Ölçekleme teknikleri*. ÖSYM Yayınları. [Scaling Techniques]. Student Selection and Placement Center (SSPC), Ankara.

Yalçın, S. & Şengül Avşar, A. (2014) Eğitim fakültesi meslek bilgisi derslerinin sıralama yargıları kanunuyla ölçeklenmesi [Scaling pedagogy courses faculty of education with rank-order judgments]. *Journal of Measurement and Evaluation in Education and Psychology*, 5 (2), 79-90.

Yaşar, M. (2016). Öğretmen adaylarının akademik başarısını etkilediği düşünülen özelliklerin sıralama yargıları yöntemine dayalı ölçeklenmesi. [Scaling the features considered to have affected the academic success of teacher candidates on the basis of rank-order judgement scaling technique]. *Pamukkale University, Journal of education*, 40, 274-288.

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

The Development of a General Disaster Preparedness Belief Scale Using the Health Belief Model as a Theoretical Framework

Ebru Inal, Kerim Hakan Altintas, Nuri Dogan

To cite this article: Inal, E., Altintas, K. H., & Dogan, N. (2018). The Development of a General Disaster Preparedness Belief Scale Using the Health Belief Model as a Theoretical Framework. *International Journal of Assessment Tools in Education*, 5(1), 146-158. DOI: 10.21449/ijate.366825

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 146-158

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

The Development of a General Disaster Preparedness Belief Scale Using the Health Belief Model as a Theoretical Framework

Ebru Inal[*1], Kerim Hakan Altintas[2], Nuri Dogan[3]

[1]Yalova University, Yalova Vocational School, Civil Defence and Firefighting Department, Yalova, Turkey
[2]Hacettepe University, Faculty of Medicine, Department of Public Health, Ankara, Turkey
[3]Hacettepe University, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Ankara, Turkey

Abstract: The Health Belief Model (HBM) is one of the oldest and most recognized conceptual framework of health behavior and can be applied to disaster preparedness efforts which focus predominantly on human behavior. The study aims to develop and test the psychometric properties of the General Disaster Preparedness Belief (GDPB) scale based on the HBM. A study group of 286 academic and administrative staff working in a Turkish University located in the city of Yalova completed a GDPB scale instrument containing 60 items. Exploratory Factor Analyses (EFA) was used for the construct validity of scale. Item analysis was assessed using item–total correlations and Cronbach's alpha coefficients. The EFA extracted six factors that jointly accounted for 59.2% of variance observed namely; Self efficacy (8 items), Cues to action (5 items), perceived susceptibility (6 items), perceived barriers (6 items), perceived benefits (3 items) and perceived severity (3 items). Cronbach's alpha coefficient for the subscales ranged from 0.90 to 0.74. The GDPB scale based on the HBM was found to be a valid and reliable tool. Findings from this study can be used to guide intervention aimed at informing and educating people about disaster preparedness.

ARTICLE HISTORY

Received: 25 September 2017
Revised: 22 November 2017
Accepted: 12 December 2017

KEYWORDS

Disaster preparedness beliefs,
Health Belief Model,
Reliability and validity,
Scale development

1. INTRODUCTION

Disasters could be natural or man-made emergency events which have negative economic and social consequences for the affected population (Donahue & Joyce, 2001). The 20th century had witnessed an increase in disaster losses, and this has continued in an upward trend in the current century (Guha Sapir, Hoyois & Below, 2013; IFRC, RCS, 2013). Turkey is under the danger of natural disaster as a result of its position which is on a young and active mountain zone called Alp-Himalaya based on a geological point of view (Ersoy & Kocak, 2015). Turkey has also witnessed its own share of disasters ranging from earthquake, landslide, and floods (Gokce, Ozden & Demir, 2008). However, in Turkey, the earthquake disasters that occurred in August 17, 1999 in Kocaeli and November 12, 1999 in Duzce were among the most devastating disasters. The 1999 Kocaeli earthquake alone left 17,000 people dead,

[*]Corresponding Author E-mail: ebruinal34@hotmail.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.366825**

200,000 homeless, and resulted in a fiscal cost of some US$2.2 billion (Ersoy & Kocak, 2015). To reduce vulnerability and increase mitigation level to disasters in Turkey and other countries, there is a need for effective disaster preparedness.

Disaster and emergency preparedness efforts focus predominantly on human behaviors derived from diverse factors that range from people's risk perception to lessons from direct and indirect past experiences of disaster events and emergencies (Ejeta, Ardalan & Paton, 2015). According to literatures, theories could be used to explain the structural and psychological determinants of behaviour as well as guide the development and refinement of health promotion and education (Painter, Borba, Hynes, Mays & Glanz, 2008).

The Health Belief Model (HBM) is one of the oldest and most widely used models in which theory has been adapted from the behavioural sciences to health problems (Glanz, Rimer & Lewis, 2002; Orji, Vassileva & Mandryk, 2012). The HBM describes the decision-making process that individuals employ when adopting a health protective behavior (Sharma & Romas, 2008). Though the use of the HBM is very versatile (Teitler-Regev, Shahrabani & Benzion, 2011; Akompab et al., 2013; Guvenc, Aygul, Acıkel, 2011; O'Connell, Price, Roberts, Jurs, McKinley, 1985), it can be beneficial when discussing disaster preparedness, because it can be applied to encourage individuals to change a potentially detrimental behavior. In the current study, behavior is seen as an intentional or unintentional lack of preparedness for imminent occurrence of disaster. In the HBM, disaster preparedness will depend on the following predictors: perceived susceptibility of experiencing a disaster, perceived severity of disaster, benefits of being prepared for a disaster, perceived barriers to being prepared, cues to action for disaster preparedness and individual's belief in their own ability to deal with a disaster (Glanz, Rimer, Lewis, 2002; Rosenstock, 1966; Rosenstock, Strecher & Becker, 1988).

Past studies have been carried out with regards to earthquake preparedness at the individual level, some of these studies have used brief measures with 10 items and below (Farley, 1993; Showalter, 1993; McClure, Walkey, Allen, 1999) to assess earthquake preparation, whereas some other studies have used longer measures between 12 and 27 items to examine more than one category of disaster preparedness such as survival, planning, and hazard mitigation (Mileti, Fitzpatrick, 1992; Mulilis, Duval, Lippa, 1990; Spittal et al., 2006). However, there are limited research work with regards general disaster preparedness with some few published researches on specific disaster preparedness topics such as heat waves and climate change; collaborative activities between non-professional disaster volunteers and victims of earthquake disasters; climate change and climate variability; as well as preparation of health care workers for disasters (Haraoka et al., 2012; Akompab et al., 2013; Semenza, Ploubidis, George, 2011; Ogedegbe, 2012). In addition, a review of the literatures revealed that there is a paucity of published papers that attempts to develop and validate instruments aimed at measuring General Disaster Preparedness Belief (GDPB) using the health behavior models as a theoretical framework. This study aims to identify scale items that have a consistent factor structure for measuring GDPB using the HBM as a framework. The findings of this study should guide the development of behaviour change programs as it relates to general disaster preparedness. The scale could also be an important tool in improving the motivation for adaptation and mitigation to related general disaster preparedness risks as well as promoting behaviour change strategies for general disaster preparedness.

2. METHOD

2.1. Study setting

The scale development study was conducted in the city of Yalova, Turkey among Yalova University staffs.

2.2. Instrumentation

An initial 78 instrument items were developed by the researchers based on current literature reviews. The initial items pool was subjected to further review by a panel of nine content experts who had expertise in the field of disaster management (6 individuals), instrument development, health education (2 individuals) and Turkish language (1 individual). The content validity index cut off was set at 0.80 which refers to the proportion of experts who rate an item as a 3 or 4 using a 4-point ordinal rating scale ranging from "1" (not relevant) to "4" (very relevant) (Davis, 1992). The experts had high harmony in terms of the content validity and no new items were recommended, on the other hand, on the basis of the content validity, the items were reduced to 60 items and then administered in a pilot study to a convenience sample of 21 individuals in order to ascertain the degree of difficulty and clarity of the items. The final scale consisted of 60 items according to six subscales namely; Susceptibility, Severity, Barriers, Benefits, Cues to action, and Self-efficacy.

2.3. Data collection

To ensure a conceptually clear factor structure for analysis, existing literature suggest a minimum sample of 3-6 respondent per item (Cattell, 1978). The desired minimum sample size for factor analysis in this study was determined to be 180 (Guilford, 1954; Gorsuch, 1983; Kline, 1979; Akgül, 1997; Tabachnick, Fidell, 2007). The scales were self-administered and were administered between April and July, 2014. The inclusion criterion for this study was willingness to participate in the study and being a staff member of Yalova University. After removal of participants with missing item response, our sample consisted of a total of 286 academic and administrative staff who had usable data for the study. Participants with missing data were removed from the study as they did not answer most of the items. During data collection, the main priority was to achieve a sufficient sample size for the analysis. The sample size of 286 participants included in the study exceeded the minimum threshold of 180 required for the study. Also, during data collection, a balance in the number of academic and administrative staff as study participants was taken into consideration however, academic staff were more willing as compared to administrative staff to participate in the study, thus, most of participants were academic staff.

2.4. Study Group

The mean age of the 286 participants was 32.8 years (±5.4 years). 69.7% of respondents were academic staff whereas 30.3% were administrative staff. A larger proportion of respondents were males (63.3%). Approximately 53% of respondents were currently married and half of the participants had a monthly salary of 2.500-2.999 Turkish lira (TL) (854 $-1025 $).

2.5. Ethics

Permission to conduct the study was obtained from relevant authorities in Yalova. Ethical approval was also taken from the Ethical Committee of Hacettepe University. All university staff who participated in the study were given informed consent letters and informed about the purpose of the study. Furthermore, they were also instructed that withdrawal from the study was optional at any time.

2.6. Measures

Respondents completed sub-scales assessing "susceptibility (9 items)", "severity (5 items)", "benefits (5 items)", "barriers (19 items)", "Cue to action (7 items)" and "self-efficacy (15 items)". All items were scored on a five point Likert scale from 1 (strongly disagree) to 5 (strongly agree). All sub-scales measured General Disaster Preparedness Belief and where negatively worded statements were used, the scores on the items were reverse-scored so that a

higher score represented more positive belief. A total scale score was computed by summing up all the 6 subscales (Self Efficacy + Cues to action + Perceived susceptibility + Perceived low barrier (items were reverse scaled) + Perceived benefits + Perceived severity).

2.7. Statistical analysis

To determine the validity of our scale we conducted an Exploratory Factor Analysis (EFA) with varimax rotation that maximizes variance explained by factors using SPSS 19. This analysis was conducted on the basis of polychoric correlation matrix. If the model includes variables that are ordinal a factor analysis can be performed using a polychoric correlation matrix. The polychoric correlation is a technique for estimating the correlation between two ordinal scales' scores (Olsson, 1979).

The Kaiser-Meyer-Olkin (KMO) was used to assess sampling adequacy while Bartlett sphericity test was used to test whether the data have a multivariate normal distribution. The factor retention criterion included the following: diagonals of the anti-image correlation matrix over 0.5, communalities above 0.3, loadings equal to or greater than 0.40, more than three items per factor, and cross-loading analysis (Fabrigar, Wegener, MacCallum, Strahan, 1999; Child, 2006), in addition, items were permitted to load only on the construct they theoretically represented as the scale was theory driven. If these constraints were not met, each item was examined individually and items were removed one at a time to ensure appropriate removal. The distribution of the total scale and sub-scale scores were described by calculating score range, mean, standard deviation, skewness and kurtosis as well as the floor and ceiling effects. Floor and ceiling effects were considered present if more than 15% of respondents achieved the highest or lowest possible score, respectively (McHorney & Tarlov, 1995). The item-total subscale correlations were assessed to determine the discrimination power of the items. While these correlations were calculated, score of calculated item was removed from total score to prevent heightening the relationship between items and scale. Reliability was assessed using Cronbach's alpha coefficients while stratified alpha was calculated for total scale score. Subscale/total scale score intercorrelations were assessed using Pearson correlation. In addition, test-retest reliability was evaluated for the study. The three week test-retest reliability coefficient for scale on the 60 item was .73. An intraclass correlation coefficient of ≥0.70 was considered as evidence of measurement stability.

3. RESULTS

3.1. Exploratory Factor Analysis

EFA using principal component analysis was used to extract factors. Various rotated analysis was computed which lead to the removal of 29 items and retention of 31 items. During several steps, a total of 20 items were eliminated because they did not contribute to a simple factor structure and failed to meet a minimum criterion of having a primary factor loading of .4 or above. In addition, 9 items had similar factor loadings. The factor loading was approved if it was at least 0.1 higher than the next higher loading (Büyüköztürk, 2002) so the 9 items were inappropriate so were eliminated.

In the final rotated analysis, the KMO value of the data was found to be 0.85. The Bartlett's test was significant (chi square =4351;00 df=496; p=<0.0001). The diagonals of the anti-image correlation matrix though not shown were all over 0.5 supporting the inclusion of each item in the factor analysis. In addition, the communalities were all above 0.3.

The factor analysis extracted six factors that jointly accounted for 59.2% of variance observed. The first factor (self-efficacy) assessed individuals' belief in their own ability to deal with a disaster/emergency and accounted for the highest proportion of scale variance (26.2%) with loadings ranging from 0.781 and 0.676. The second factor (susceptibility) addressed

perceived risk of experiencing an emergency or disaster, and this accounted for 9.8% of variance and the loading ranged from 0.735 to 0.491. The third factor (cue to action) related to events, people, or other exposures that could influence disaster preparedness behaviour, accounted for 8.0% of the variance with loading ranging from 0.795 to 0.629, while the fourth factor (barrier) related to perceived obstacles that could hinder disaster preparedness, this factor accounted for 5.8% of the variance and had a loading of 0.789 to 0.426. The fifth factor (benefit), addressed belief about the benefit of disaster preparedness and accounted for 5.6% of the variance and had a loading range of 0.794 to 0.732 while the sixth factor (severity) relating to fear of disaster and belief about the consequences of disaster accounted for 4.3% of the variance and had a loading range of 0.773 to 0.722 (Table 1).

Table 1. Rotated Factor Solution of General Disaster Preparedness Belief (n = 286)

Items (n = 31)	**Self-efficacy**	**Susceptibility**	**Cues to action**	**Low barrier**	**Benefit**	**Severity**	**Communalities**
eff1	0.781	*	*	*	*	*	0.634
eff2	0.778	*	*	*	*	*	0.715
eff3	0.763	*	*	*	*	*	0.748
eff4	0.745	*	*	*	*	*	0.636
eff5	0.710	*	*	*	*	*	0.546
eff6	0.707	*	*	*	*	*	0.542
eff7	0.703	*	*	*	*	*	0.612
eff8	0.676	*	*	*	*	*	0.637
sus1	*	0.735	*	*	*	*	0.612
sus2	*	0.729	*	*	*	*	0.606
sus3	*	0.687	*	*	*	*	0.556
sus4	*	0.664	*	*	*	*	0.513
sus5	*	0.521	*	*	*	*	0.374
sus6	*	0.491	*	*	*	*	0.356
cue1	*	*	0.795	*	*	*	0.732
cue2	*	*	0.786	*	*	*	0.658
cue3	*	*	0.769	*	*	*	0.620
cue4	*	*	0.762	*	*	*	0.628
cue5	*	*	0.629	*	*	*	0.537
bar1	*	*	*	0.789	*	*	0.686
bar2	*	*	*	0.786	*	*	0.738
bar3	*	*	*	0.562	*	*	0.588
bar4	*	*	*	0.515	*	*	0.384
bar5	*	*	*	0.450	*	*	0.447
bar6	*	*	*	0.426	*	*	0.379
ben1	*	*	*	*	0.794	*	0.738
ben2	*	*	*	*	0.776	*	0.718
ben3	*	*	*	*	0.732	*	0.655
sev1	*	*	*	*	*	0.773	0.667
sev2	*	*	*	*	*	0.760	0.632
sev3	*	*	*	*	*	0.722	0.617
Eigenvalues	8.133	3.039	2.486	1.791	1.730	1.333	
% of variance	26.24	9.80	8.02	5.78	5.58	4.30	

Not: R=Reverse scored, Asterisk (*) is less than 0.40.

3.2. Descriptive statistics for items, internal consistency and descriptive statistics for subscales and total scale

Ceiling and floor effects were negligible for most of the 31 items. Ceiling effects were observed for 3 items in the susceptibility subscale, 3 items in the benefit subscale and for 3 items in the severity subscale. Whereas floor effect was observed for 1 item in the cue to action subscale and 2 items in the susceptibility subscale. Overall, there was no evidence that there was a systematic response pattern which could be interpreted as a sign of the participants' reflection of their thoughts (Appendix 1).

The internal consistency of the total scale and subscales all exceeded 0.70 showing that the scale is reliable, the internal consistency for subscales ranged from 0.74 to 0.90. For the total scale, the stratified alpha was 0.93. The mean score for self-efficacy subscale was 24.7±6.4 and for susceptibility subscale was 22.3±3.8. Ceiling effect was observed for the severity subscale. The mean score for the total scale was 102.3±15.3 (Table 2).

Table 2. Item Total Subscale Correlation, Reliability Coefficients and Descriptive Statistics for Sub Scales and Total Scale

Subscale	No of scale item	Item total subscale correlation	Cronbach alpha	Mean	SD	Skewness	Kurtosis	Range: Observed (Possible)	Floor %	Ceiling %
Self-Efficacy	8	0.69-0.84	0.90	24.69	6,35	-0.20	-0.95	9-38 (8-40)	0	0
Cues to Action	5	0.70-0.84	0.84	13.21	4.03	0.20	-0.68	5-24 (5-25)	1.0	0
Perceived Susceptibility	6	0.59-0.73	0.76	22.31	3,78	-0.48	0.03	11-30 (6-30)	0	2.4
Perceived low Barriers	6	0.57-0.78	0.75	18,58	4,07	0.03	-0.98	10-28 (6-30)	0	0
Perceived Benefits	3	0.83-0.87	0.80	11,93	1.95	-1.13	2.75	4-15 (3-15)	0	9.8
Perceived Severity	3	0.80-0.83	0.74	11,53	2,45	-0.80	0.81	3-15 (3-15)	0.7	15.7
Total scale score (stratified alpha)	31	0.38-0.71*	(0.93)	102.27	15.28	-0.28	-0.22	62-138 (31-155)	0	0

Not: *Item total correlation

3.3. Item-total subscale correlations and total item correlations

The item-total subscale correlations were as follows; Self-Efficacy ranged from 0.69 to 0.84; Cue to Action ranged from 0.70-0.84; Perceived Susceptibility ranged from 0.59-0.78; Perceived low barriers ranged from 0.57-0.78; Perceived Benefits ranged from 0.83-0.87 whereas Perceived Severity ranged from 0.80-0.83. The total item correlation for the total scale score and items ranged from 0.38-0.71 (Table 2).

3.4. Subscale/Total Scale score intercorrelations

The six derived subscales had an intercorrelation range between subscales of 0.22 to 0.46 ($p<0.01$), the correlation were weak or moderate between the subscales highlighting the unique contributions of each subscale in understanding general disaster preparedness beliefs. The total scale score correlations with the 6 subscales all exceeded the .50 level, 5 of the 6 coefficients

exceeded the .60 level, and 2 of the 6 exceeded .70. All correlations were less than 0.01 level of probability, indicating that even the weakest of the relationships was nonetheless significant. The fact that the correlation coefficients were significant between the 6 subscales and the total scale score could be taken as evidence for summing up all the 6 subscales and for using the total test scores (Table 3).

Table 3. Subscale/Total Scale Intercorrelations

	Self-Efficacy	Cues to Action	Perceived Susceptibility	Perceived Benefits	Perceived low Barriers	Perceived Severity	Total scale score
Self-Efficacy	1.000						
Cues to Action	0.291**	1.000					
Perceived Susceptibility	0.258**	0.319**	1.000				
Perceived Benefits	0.364**	0.236**	0.461**	1.000			
Perceived Barriers	0.453**	0.364**	0.381**	0.412**	1.000		
Perceived Severity	0.368**	0.149**	0.217**	0.286**	0.243**	1.000	
Total scale score	0.783**	0.615**	0.634**	0.610**	0.737**	0.507**	1.000

**p<0.01

4. DISCUSSION

In Disaster Risk Reduction, disaster preparedness is seen as one of the basic components. Also, effective preparedness reduces vulnerability, increases mitigation level, enables timely and effective response to a disaster event, shortens the recovery period from a disaster, and increases community resilience (Guha-Sapir, Hoyois & Below, 2013; Gregory et al., 2006). According to previous studies, the determinant of disaster preparedness behaviours include: risk perception (Armaş & Avram, 2008), preparedness perception (Mulilis & Duval, 1995), self-efficacy (McClure, Walkey & Allen, 1999), community participation (Paton, 2006) available resources and demographics (Mileti, Darlington, 1995; Najafi, Ardalan, Akbarisari, Noorbala & Jabbari, 2015).

The use of the HBM can encourage individuals to promote positive disaster preparedness habits. Accordingly, if disaster is perceived as a health threat, then the components of the HBM might be able to predict preparedness behavior. It is believed that beliefs might influence behaviour (Fabrigar et al., 2006). There are studies showing that differences in household preparedness behaviors were correlated with beliefs about preparedness (Thomas, Leander-Griffith, Harp, Cioffi, 2015; Becker et al., 2013). The HBM predicts that, "if individuals regard themselves as susceptible to a condition, believe that condition would have potentially serious consequences, believe that a course of action available to them would be beneficial in reducing either their susceptibility to or severity of the condition, and believe the anticipated benefits of taking action outweigh the barriers to (or costs of) action, they are likely to take action that they believe will reduce their risks'' (Glanz, Rimer, Viswanath, 2008). Previous studies have applied the HBM to study disaster preparedness, for instance, disease outbreak preparedness (Teitler-Regev, Shahrabani & Benzion, 2011), and preparedness for climate change and heat waves (Akompab, Bi, Williams, Grant, Walker & Augoustinos). However, in the literatures there are no studies to the best of our knowledge that have developed and evaluated a scale for GDP using the HBM as a theoretical frame work. This study attempted to evaluate a newly

developed theory driven instrument for assessing GDPB using the HBM as a framework. The study followed an established scale development process such as current literature review for the selection of items, content validity, pre-testing, scale administration and EFA.

The content validity of the items were found to be acceptable, and the EFA was able to accounted for 59.2% of the variance observed. The EFA is suitable for use on Likert-type of scale and extracted six factors measuring the following; individuals' belief in their own ability to deal with a disaster, perceived susceptibility of experiencing a disaster, perceived severity of disaster, benefits of being prepared for a disaster, perceived barriers to being prepared, and cues to action for disaster preparedness. The KMO value of the data was meritorious and above the recommended value of 0.60 (Fabrigar, Wegener, MacCallum & Strahan, 1999). The communalities confirmed that each item shared some common variance with other items (Child, 2006). Skewness and kurtosis values of each subscale were acceptable as recommended by Kline who suggest that skewness values should be lower than 3 and kurtosis values should be lower than 10 (Kline, 1998). The subscale internal consistency as estimated by Cronbach's alpha was high which in turn suggest that the items in each scale were homogeneous.

The study is not without some limitations, the participants came from a groups that had a higher than average educational and socioeconomic status, for instance, based on comparison of demographic characteristics between our study respondents and the general population, our study participants were comparatively younger males and consisted of academic and administrative staff working in a government university and earning a more or less adequate incomes. In addition, we were limited to EFA as our sample size was not large enough to split the sample into two split - half samples which would have permitted us to conduct EFA analysis on one half of the sample and Confirmatory Factor Analysis on the other half of the sample. Also, there is a need for a more detailed testing before the utility this scale can be firmly established, for example, validity and reliability could be performed in other groups using a larger sample and the scale verified by using a confirmatory factor analysis to determine the utility of the scale.

5. CONCLUSION

The result indicate that the 31 items model is a reliable and valid instrument for measuring GDPB, furthermore, the study has been able to demonstrate the application of the test and it would be interesting to applicate it in future research. Knowledge gained from this study can be used to guide intervention aimed at informing and educating people about disaster preparedness.

6. REFERENCES

Akgül, A. (1997). Tıbbi araştırmalarda istatistiksel analiz teknikleri "SPSS uygulamaları. (2. Baskı). Ankara. Emek Ofset.

Akompab, D.K., Bi, P., Williams, S., Grant, J., Walker, I.A., & Augoustinos, M. (2013). Heat waves and climate change: Applying the Health Belief Model to identify predictors of risk perception and adaptive behaviors in Adelaide, Australia. *International Journal of Environmental Research and Public Health*, 10 (6), 2164-2184.

Armaş, I., & Avram, E. (2008). Patterns and trends in the perception of seismic risk. Case study: Bucharest Municipality/Romania. *Natural Hazards*, 44(1), 147-61.

Becker, J., Paton, D., Johnston, D., et al. (2013). Salient beliefs about earthquake hazards and household preparedness. *Risk Anal*, 33, 1710-27.

Buyukozturk, Sener. *Sosyal bilimler için veri analizi el kitabı istatistik, araştırma deseni SPSS uygulamaları ve yorum*, Pegem Yayınları, Pegem Yayıncılık, Ankara, 2002.

Cattell, R.B. (1978). *The scientific use of factor analysis*. New York. Plenum.

Child, D. (2006). *The essentials of factor analysis*. Continuum, London.

Davis, L.L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, 5 (4). 194-197.

Donahue, A., & Joyce, P. (2001). A framework for analyzing emergency management with an application to federal budgeting. *Public Administration Review*, 61(6), 728-740.

Ejeta, L.T, Ardalan, A., & Paton, D. (2015). Application of behavioral theories to disaster and emergency health preparedness. A systematic review. *PLOS Currents Disasters*, 7, 2015.

Ersoy, S., & Kocak, A. (2015). Disasters and earthquake preparedness of children and schools in Istanbul, Turkey. *Geomatics, Natural Hazards and Risks*, 7(4). 1307-1336.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4 (3). 272-299.

Fabrigar, L., Petty, R., Smith, S., et al. (2006). Understanding knowledge effects on attitude-behavior consistency: the role of relevance, complexity, and amount of knowledge. *J Pers Soc Psychol*, 90 556-77.

Farley, J. E., Barlow, H. D., Finklestein, M. S., and Riley, L. (1993). Earthquake hysteria, before and after: a survey and follow-up on public response to the browning forecast. *Int. J. Mass Emergencies Disasters*. 11, 305-322.

Glanz, K., Rimer, B.K., Viswanath, K., (eds). (2008). *Health behavior and health education: theory, research, and practice.* John Wiley & Sons.

Glanz, K., Rimer, B.K., Lewis, F.M. (2002). *Health behavior and health education theory, research and practice*. San Fransisco: Wiley & Sons.

Gokce, O., Ozden, S., & Demir, A. (2008). The statistical and spatial distribution of disasters in Turkey Disaster information inventory Ankara. Turkish Ministry of Public Works and Settlement, Disaster Research and Assessment Department (pp. 118).

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Erlbaum.

Gregory, R.C, Philip, D.A, Erik, A.D.H., Robert, G.D., et al. (2006). *Disaster medicine*. U.S.A. Mosby Elsevier, pp.29.

Guha-Sapir, D., Hoyois, Ph., & Below, R. (2013). Annual disaster statistical review 2012: The numbers and trends. Brussels: CRED.

Guilford, J. P. (1954). Psychometric methods. New York: McGraw-Hill.

Guvenc, G., Aygul, A., & Acıkel, C.H. (2011). Health belief model scale for cervical cancer and Pap smear test: psychometric testing. *Journal of advanced nursing*, 67(2), 428-437.

Haraoka, T., Ojima, T., Murata, C., Hayasaka, S. (2012). Factors influencing collaborative activities between non-professional disaster volunteers and victims of earthquake disasters. *Plos One*. 7(10):e47203. doi:10.1371/journal.pone.0047203.

International Federation of Red Cross (IFRC) & Red Crescent Societies (RCS). (2013). World Disaster Report, Focus on technology and the future humanitarian action. Geneva.

Kline, P. (1979). *Psychometrics and psychology*. London: Acaderric Press.

Kline, R.B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.

McClure, J., Walkey, F., & Allen, M. (1999). When earthquake damage is seen as preventable: Attributions, locus of control and attitudes to risk, applied psychology. *Int. Rev*. 48(2): 239-56.

McHorney, C.A., & Tarlov, A.R. (1995). Individual-patient monitoring in clinical practice: are available health status surveys adequate*? Quality of Life Research*, 4 (4): 293-307.

Mileti, D., Fitzpatrick, C. (1992). The causal sequence of risk communication in the park field earthquake prediction experiment, *Risk Anal*. 12: 393-400.

Mileti, D.S., & Darlington, J. (1995). Societal response to revised earthquake probabilities in the San Francisco Bay area. *International Journal of Mass Emergencies and Disasters*, 13(2), 119-45.

Mulilis, J. P., Duval, T. S., Lippa, R. (1990). The effects of a large destructive local earthquake on earthquake preparedness as sssessed by the earthquake preparedness scale. *Nat. Hazards*. 3, 357-371.

Mulilis, J.P., & Duval, T.S. (1995). Negative threat appeals and earthquake preparedness: A person relative to event (PrE) model of coping with threat. *Journal of Applied Social Psychology*, 25(15), 1319-39.

Najafi, M., Ardalan, A., Akbarisari, A., Noorbala, A.A., & Jabbari, H. (2015). Demographic determinants of disaster preparedness behaviors amongst Tehran inhabitants, Iran. *PLOS Currents Disasters*, 7.

O'Connell, J.K., Price, J.H., Roberts, S.M., Jurs, S.G., & McKinley, R. (1985). Utilizing the Health Belief Model to predict dieting and exercising behavior of obese and nonobese adolescents. *Health education quarterly*, 12 (4), 343-351.

Ogedegbe, C., Nyirenda, T., Delmoro, G., Yamin. E., Feldman, J. (2012). Health care workers and disaster preparedness: Barriers to and facilitators of willingness to respond. *International Journal of Emergency Medicine*. s:29. Erişim Tarihi:10.07.2016. http://www.intjem.com/content/5/1/29.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), 443-460.

Orji, R., Vassileva, J., & Mandryk, R. (2012). Towards an effective health interventions design: an extension of the health belief model. *Online journal of public health informatics*, 4(3), ojphi.v4i3.4321.

Painter, J.E., Borba, C.P., Hynes, M., Mays, D., & Glanz, K. (2008). The use of theory in health behavior research from 2000 to 2005: a systematic review. *Annals of Behavioral Medicine*, 35 (3), 358-362.

Paton, D. (2006). Disaster resilience: Integrating individual, community, institutional and environmental perspectives. Disaster resilience: An integrated approach, pp. 306-19.

Rosenstock, I.M. (1966). Why people use health services. *Milbank Mem Fund Q*. 44, 94-127.

Rosenstock, I.M, Strecher, V.J., & Becker, M.H. (1988). Social Learning Theory and the Health Belief Model. *Health Education Quarterly*, 15 (2), 175-183.

Semenza, J., Ploubidis, G., George, L. (2011). Climate change and climate variability: Personal motivation for adaptation and mitigation. *Environmental Health*. 10:46. Retrieved from (14.07.2016) http://www.ehjournal.net/content/10/1/46.

Sharma, M., Romas, J.A. (2008). *Theoretical foundations of health education and health promotion*. Sudbury, MA: Jones and Bartlett Publishers.

Showalter, P. S. (1993). Prognostication of doom: an earthquake prediction's effect on Foursmall Communities. *Int. J. Mass Emergencies Disasters.* 11, 279-292.

Spittal, J W., Walkey, H F., McClure, J., Sıegert, J R., Ballantyne, E K. (2006). The earthquake readiness scale: The development of a valid and reliable unifactorial measure. *Natural Hazards*. 39, 15-29. DOI 10.1007/s11069-005-2369-9.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Allyn and Bacon.

Teitler-Regev, S., Shahrabani, S., & Benzion, U. (2011). Factors affecting intention among students to be vaccinated against A/H1N1 Influenza: A health belief model approach. *Advances in Preventive Medicine*.

Thomas, T. N., Leander-Griffith, M., Harp, V., & Cioffi, J. P. (2015). Influences of preparedness knowledge and beliefs on household disaster preparedness. *MMWR Morb Mortal Wkly Rep*, *64*(35), 965-971.

Table Appendix 1. Item Responses to Statements on General Disaster Preparedness Belief

		Percentage (%)				
		SA	A	U	D	SD
eff1	I can not create an emergency /disasters evacuation plan with the people who live around my neighbourhood (R).	9.4	33.2	21.7	32.5	3.1
eff2	I can do basic first aid.	3.1	32.5	21.0	35.0	8.4
eff3	I can specify the hazards which can cause a fire.	6.3	38.8	18.5	31.5	4.9
eff4	I can not conduct search and rescue even at the basic level (R)	4.9	36.0	21.3	35.0	2.8
eff5	I can fix the furniture that need to be fixed at home.	5.9	55.6	18.5	18.5	1.4
eff6	After an emergency situation/disaster, I can access the necessary services needed for psychological support.	9.4	39.9	28.3	18.2	4.2
eff7	I can not use a fire extinguisher (R).	8.7	29.4	21.0	33.2	7.7
eff8	I can determine a safe place at home/in the building to stay during an earthquake.	2.8	40.6	25.2	27.3	4.2
cue1	The policies on emergency situation/disaster encourage me to be prepared for emergency situations/disasters.	3.1	22.0	21.7	39.2	14.0
cue2	My friends enlighten me about the necessity of making individual preparations for emergency situations/disasters.	0.3	12.9	22.0	46.2	18.5
cue3	Booklets, newspapers, brochures do not inform me enough (R).	8.7	42.0	18.5	26.2	4.5
cue4	The people to whose opinion I pay much importance to guide me on the subject of emergency /disaster preparedness.	3.1	26.9	20.3	38.5	11.2
cue5	My family members do not inform me about the necessity of making individual preparations for emergency situations/disasters (R).	7.0	37.8	27.3	22.7	5.2
sus1	I do not attach importance to preparing emergency/disaster kit for emergency situations/disasters preparation (R).	0.7	14.7	21.7	44.4	18.5
sus2	I take into consideration that I may experience an emergency situation/a disaster at some point in my life	15.4	59.4	12.6	12.6	0.0
sus3	It is important for me to enhance building durability in the case of emergency situations/disasters preparation.	36.0	49.3	10.8	3.8	0.0
sus4	My possibility of experiencing an emergency situation/a disaster is very high in the next couple of years.	15.4	49.7	23.1	7.7	4.2
sus5	I find it unnecessary to fix the furniture that need to be fixed at home(R).	0.0	6.3	13.3	56.3	24.1
sus6	I do not talk about necessary emergency contact numbers during emergency situations/disasters in my neighbourhood (R).	7.7	42.3	12.6	25.5	11.9
bar1	It takes too much time of mine to make individual preparations for emergency situations/disasters.(R)	4.2	42.3	15.0	33.2	5.2
bar2	I have responsibilities more important than making preparations for emergency situations/disasters.(R)	0.0	38.8	11.2	45.1	4.9
bar3	I do not have enough information on individual emergency/disaster preparedness (R).	8.7	54.2	17.1	16.4	3.5
bar4	I do not have enough money to make preparations for emergency situations/disasters.(R)	0.0	17.5	15.7	61.2	5.6
bar5	If it is my destiny to die as a result of emergency situations/disasters, I will die (R).	3.1	38.1	15.7	31.1	11.9
bar6	I find it difficult to understand the family disaster plan(R).	3.5	22.7	20.6	44.4	8.7
ben1	My making individual preparations for emergency situations/disasters will also save my family members.	19.9	63.3	10.5	3.1	3.1
ben2	Making preparations for emergency situations/disasters is helpful for my needs during emergency situations/disasters	19.6	61.9	15.7	2.1	0.7
ben3	Making individual preparations for emergency situations/disasters may decrease the risk of death after emergency situations/disasters.	21.7	65.0	7.7	4.9	0.7
sev1	An emergency situation/a disaster experience would not change my life (R).	4.9	6.3	5.9	60.5	22.4
sev2	I am afraid of dying as a result of emergency situations/disasters.	23.8	52.8	9.1	8.4	5.9
sev3	The idea of disasters scares me	19.6	59.1	11.9	4.9	4.5

SA = 5=Strongly Agree (SA), 4 = Agree (A), 3 = Uncertain (U), 2= Disagree (D), 1= Strongly Disagree (SD). R=Reverse coded

Table A2. Item Responses to Statements on General Disaster Preparedness Belief (Turkish Version)

eff1	Mahallemde yaşayanlarla birlikte Acil durumlar/Afetler ile ilgili tahliye planı oluşturamam.
eff2	Temel ilk yardım uygulayabilirim.
eff3	Yangın çıkmasına neden olacak tehlikeleri belirleyebilirim.
eff4	Basit düzeyde olsa dahi arama-kurtarma yapamam.
eff5	Evde sabitlenmesi gereken eşyaları sabitleyebilirim.
eff6	Acil durum/afet sonrası ihtiyacım olursa psikolojik destek almak için gerekli hizmete erişebilirim.
eff7	Yangın söndürme cihazını kullanamam.
eff8	Depremden korunmak için yaşadığım evde/binada güvenli yer belirleyebilirim.
cue1	Acil durum/Afet konusundaki politikalar beni Acil Durumlar/Afetler konusunda hazırlıklı olmaya teşvik ederler.
cue2	Arkadaşlarım Acil durumlara/Afetlere bireysel hazırlık yapmanın gerekliliği konusunda beni aydınlatırlar.
cue3	Kitapçıklar, gazeteler, broşürler beni yeterince bilgilendirmezler.
cue4	Fikirlerine önem verdiğim insanlar acil durumlara/afetlere hazırlıklı olma konusunda beni yönlendirirler.
cue5	Aile üyelerim Acil durumlara/Afetlere bireysel hazırlık yapmanın gerekliliği konusunda beni bilgilendirmezler.
sus1	Acil durumlara/Afetlere hazırlıkta acil durum/afet çantası hazırlamayı önemsemem.
sus2	Yaşamımın herhangi bir döneminde Acil durum/Afet yaşayacağımı göz önünde bulundururum.
sus3	Acil durumlara/Afetlere hazırlıkta bina dayanıklılığını artırmak benim için önemlidir.
sus4	Önümüzdeki birkaç yıl içinde Acil durum/Afet yaşama ihtimalim çok yüksektir.
sus5	Evdeki sabitlenebilecek eşyaları sabitlemeyi gereksiz buluyorum.
sus6	Yakın çevrem ile acil durumlarda/afetlerde gerekli acil iletişim numaraları hakkında konuşurum.
bar1	Acil durumlara/Afetlere bireysel hazırlık yapmak çok fazla zamanımı alır.
bar2	Acil durumlara/Afetlere hazırlık yapmaktan çok daha önemli sorumluluklarım var.
bar3	Acil durumlara/Afetlere bireysel hazırlık yapmak için yeterli bilgim yok.
bar4	Acil durumlara/Afetlere hazırlık yapmak için yeterli param yok.
bar5	Kaderimde Acil durumlarda/Afetlerde ölmek varsa ölürüm
bar6	Aile için afet planının anlaşılması zordur.
ben1	Acil durumlara/Afetlere bireysel hazırlık yapmam aile bireylerimi de koruyacaktır.
ben2	Acil durumlara/Afetlere hazırlık yapmak acil durumlarda/afetlerde ihtiyaçlarıma karşılık verecektir.
ben3	Acil durumlara/Afetlere bireysel hazırlık yapmak acil durumlar/afetler sonrası ölüm riskini azaltabilir.
sev1	Acil durum/Afet yaşarsam hayatımda hiçbir şey değişmeyecek.
sev2	Acil durumlar/Afetler sonucunda ölmekten korkarım.
sev3	Acil durum/Afet yaşama ihtimalini düşünmek beni korkutur.

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Look Sir, I Drew You

Ulas Kubat

To cite this article: Kubat, U. (2018). Look Sir, I Drew You. *International Journal of Assessment Tools in Education*, 5(1), 159-175. DOI: 10.21449/ijate.370386

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 159-175

http://www.ijate.net

e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Look Sir, I Drew You

Ulas Kubat

Muğla Sitki Koçman University, Faculty of Education, Mugla, Turkey

Abstract: The purpose of this study is to try to find out how the fifth, sixth, seventh and eighth graders perceive science teachers through the pictures they have drawn. A qualitative research method was used in the research. A total of 246 students studying in 5th, 6th, 7th and 8th grade, using the appropriate sampling method, constitute the study group of the research. The students were asked to draw pictures when asked "what comes to mind when you think about a science teacher." In the research, it was found that most of the students perceive the science teacher as "human," while a few of them perceive it as a cartoon character or a famous scientist like "Einstein." The students reflected the science teacher's gender more often as female than male. While about one-third of the students drew science teachers as in the classroom, none of the students reflected the teachers in their pictures as in non-school learning environments like a museum or a science center.

ARTICLE HISTORY

Received: 04 November 2017

Revised: 11 December 2017

Accepted: 20 December 2017

KEYWORDS

Students' pictures,
science teacher,
teacher perception

1. INTRODUCTION

Teachers are one of the most important elements of the learning-teaching process. In this process, teachers, as one of the most important elements, structure the learning-teaching process and prepare a rich learning environment for the students. In the learning-teaching process, the influence of many variables such as the teacher's professional knowledge and skills, the teaching methods and strategies chosen, the use of equipment, classroom management, the physical conditions of the class, the level of readiness of the students, and the differences of the individual are very important.

In the learning-teaching process, the teacher should choose teaching strategies and methods appropriate to the purpose of the course. It would not be possible to achieve the desired result with teaching strategies and methods that are not selected according to students' achievements. Since the learning speed, readiness and motivation will differ from student to student, activities organized in the learning-teaching process should be organized in this direction. Research emphasizes the importance of instructional strategies that motivate, question, and support the student in relation to the real life of the student (Corbett & Wilson, 2002; Thompson, 2002). In other words, it is possible to create a rich learning environment only if one takes into account the needs of students.

*Corresponding Author E-mail: ulaskubat@mu.edu.tr

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.370386**

One of the important tasks of the teacher is to better identify the individual differences of the students and to better determine the needs of their students. It is unlikely that a teacher will be able to create a rich learning environment that does not adequately reveal pupils' reading and readiness levels. Individual differences in learners are always an important part of teaching. Teachers need a variety of different teaching strategies to accomodate the various needs of students (Jacobsen, Eggen, Kauchak, 2006: 284). The most important task is to develop and apply teaching methods and techniques according to different learning styles of each student. This way, students can learn in a way that both appeals to them, as well as addresses the subjects they feels they need by allowing more active participation.

It is the first duty of a teacher to properly design and use tools in the right time and place. It is important that these selected tools, besides being appropriate for the lesson's subject and purpose, must low-cost and easy to obtain. In addition, the ease of use of these tools and the level of development of students should be considered when selecting appropriate tools.

The physical conditions of classroom environment are among important factors for student success. The physical characteristics of the classroom such as width, height, color, light, cleanliness, accoustics and aesthetics, along with a seating plan are all important factors for student success in the learning-teaching process (Gökçe, 2014: 73). It is emphasized that there is a connection between educational outcomes and physical conditions of schools (Clark, 2002). The acoustic structure, color, lighting, comfort, and classroom design of educational facilities should be well considered when creating an effective learning environment (Dudek, 2000, Clark, 2002). In other words, the learning environment being well-lit, well-warmed, having comfortable seating and being suitably painted, will contribute to students' success. The physical appearance of a class should be designed to complement student activities whilst taking their needs into consideration. (Burden, 1995). The rate of student success with teachers who provide a positive, intimate, student-supportive classroom atmosphere, is higher than those whose classroom environments are negative, unpleasant, or unsupportive of students. (Moore, 2001: 53). For this reason, teachers should prepare their classes very carefully at the beginning of each school year.

It is important how a student perceives the teaching-learning process structurer that is the teacher. Drawing, painting, and three-dimensional building activities are concrete indicators of a child's emotions, thoughts, concepts, reactions and skills. Each child interperates their surroundings differently (Artut, 2002). The drawings made by children reflect their inner world (Malchiodi, C. A, 1998). These are effective ways of exploring children's thoughts, their perceptions, and their inner world (White and Gunstone, 1992). Drawings made by children of different ages are an important sign of their mental development, which is one of the best ways of expressing their emotions. (Lowenfeld & Brittain, 1987). Some of the lines, symbols and signs that children come to possess with perception are very important, and they are components that reflect the world of children plainly. Painting is also a unique and simple expression of the emotional and intellectual life of children (Artut, 2002). Therefore, children are expressing their thoughts and feelings about the pictures and events they have been experiencing and been thorugh in their lives

The fact that pictures and children are a dynamic in which they complement one another and that besides pictures being proof of how people identify themselves are a rather effective method in perceiving and representing nature. Apart from uncovering children's feelings, drawings also provide insight into their cognition, thoughts, perceptions, and judgments. (Lin, 2006). Drawings are often used in research to study the insights and perspectives of individuals. They are therefore a useful way to examine the process of teacher identity development in students (Weber & Mitchell, 1996). Children's drawings are one of the best ways of self expression. Children can freely express emotions and thoughts with colours, lines, shapes, and

images that they use. (Hsiao & Chen, 2015). In this context, we can get a lot of information about teachers from the pictures children have drawn.

The cited research focuses on the perception of "scientists" of middle school students at the age of 12-13, consisting of one experimental group and one control group pictures they drew (Gültekin, Ç., Tosun, Ö., Turgut, Ş., Örenler, Ş., Şengül, K. and Top, G., 2010). In another study, environmental perceptions of elementary school students were studied through painting (Özsoy, 2012). Analysis of the pictures drawn by the students reveals that although the new science program is student-centered, there still exists a more teacher-centered learning environment in science classes (Skoumios, MariaSavvaidou-Kambouropoulou; 2012). No research has been conducted so far to reveal students' perceptions of science teachers through drawings. Therefore, with this study, it will be possible to obtain important information about students' perceptions of science teachers during the learning process, as well as the actual teachers' teaching-learning process itself.

The aim of this research is to determine the primary school students' perception of science teachers. The research attempts to reveal primary school students' perceptions through drawings, of science teachers, their facilities, tools and teaching materials, which postures and facial expressions they use and the kinds of activities they implement. In other words, with the help of the pictures, theesearcher attempts to find out how science teachers form the learning-teaching process.

2. METHOD

A qualitative research method was used in this study. In the qualitative research, the researcher works on the events without interfering with the natural state of formation. The product of the qualitative research is usually based on a rich detailed and in-depth narrative rather than a statistical testimonial that includes a multitude of statistical test results (Johnson, Christensen, 2012). The fifth, sixth, seventh, and eighth grade students have participated in this research to reveal their perception of both the practice of science education and the learning-teaching processes. They have created an in depth and thorough examination of their science teacher through their drawings.

2.1.Working Group

The study group consists of 246 students studying in the primary schools affiliated with the Ministry of National Education of Turkey. The purposive sampling methods were used in the research. In the purposive sampling methods, the researcher forms the study group from the sample that is easiest to access (Cohen, Manion, Morrison, 2000). The purposive sampling method provides time, money and labor savings (Büyüköztürk et al., 2009).

Table 1. Distributions of Surveyed Students by Grades

Students Grades	Student Frequencies f / %
Fifth grade	81 / 32.92
Sixth grade	44 / 17.88
Seventh grade	76 / 30.89
Eighth grade	45 / 18.29
TOTAL	246 / 100.0

2.2. Data Collection

The students in the study group were directed to the question "What comes to mind when you think about science teachers?" and asked to draw a picture of it. Before the drawing, students were provided with paper, pencils, colouring pencils and oil pastels which they could choose and draw with. There was no guidance about what to draw. Students were given 45 minutes to complete their paintings. In qualitative research, visual materials such as film, video and photographs can be used as data collection tools. When such materials are used together with data collection methods such as observation, interview and document analysis, the reliability of qualitative research based on collected data in such a versatile method will increase significantly (Yıldırım, Şimşek, 2000). The data was collected during the spring of 2015-2016 period.

2.3. Analysis of Data

The "Drawing Analysis Scientist Test" (DAST) method developed by Chambers (1983) distinguishes the typical scientist image from seven main characteristic features. However, Finson and Beaver (1995) developed this criterion as the "Drawer Scientist Test-Checklist (DAST-C)", which is easily applicable to anyone. In this study, a "perception of teacher" checklist consisting of 13 categories and subcategories of the scientist drawing test and the scientist control list created by Aykaç (2012) was used.

In this research, 'Perception of Teacher Coding List' which was developed thanks to expert opinions by Aykaç (2012) has been consulted. The categories in the "Perception of Teacher Coding List" are "gender," "size," "gestures and facial expressions," "physical features," "facility," "actions taken," "object used in hands," and "objects found in class." The digitized values from the categories were obtained and tabulated by using the SPSS program, percent (%) and frequency values. Findings reached in the research are presented by interpreting the data in the tables.

3. RESULTS

The frequency data of 246 images obtained as a result of the research were analyzed using the SPSS packet program and the findings are tabulated in percentage and frequency. In the analysis of the drawings, a " perception of teacher " checklist consisting of 13 categories and subcategories was created by Aykaç (2012). The checklist used was formed in a similar manner to the scientist control list and was finalized by reffering expert opinions. All students' drawings were evaluated and interpreted according to these categories listed below. The following categories created for drawings are listed:

1. The way pupils perceive their teacher (Human, a recognized person, cartoon character, object, etc.).
2. Gender perceptions of learners about the teacher (female, male, not human, uncertain, etc.)
3. Physical appearance (in suit, white gown, tie, scarf, scattered, young, etc.)
4. Metaphores drawn as teachers (sun, book, heart, moon, star, cloud, school, flower, world, angel, etc.)
5. Dimension (There is big, there is small, realistic.)
6. Gestures and facial expessions (happy face, excited, confused, angry, sad, shy, anxious, unhappy, thoughtful etc.)
7. Physical properties (with glasses, scattered hair, clean groomed, bald, bearded, mustache, physical disabilities, remarkable wounds, etc.).

8. Place / facility/ positioning (Class, front of table, side, desk, laboratory, teacher's room, garden, ceremony, event, computer, next to the flagpole, sky, etc.)
9. Form of action (When writing on the board, talking to the students, reading the paper, reading the book, lecturing, listening, experimenting, violence against the students,
10. Objects used in hands (Ruler-stick, chalk, book, bag, paper, flower, pen, ball, bar pallet etc.)
11. Objects around you (Library, students, table, board, tree, flower, heart, etc.)
12. Layout plan (Traditional layout layout, semi-layout, layout u, set layout, free layout, etc.)
13. Objects and objects found in the class (wooden, table, row, cabinet, computer, projection device, etc.)

While the student pictures were examined, the uncollected categories were coded as "undrawn" and the drawings other than the specified categories are given under “the others” heading. Frequencies and percentages were used and interpreted when the data was evaluated.

Table 2. Students’ Perceptions of Teacher

Perceptions	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
Human	62 / 76.54	29/65.90	64 / 84.21	39 / 86.66	194 / 78.86
A Recognized Person	3 /3.70	2 / 4.54	4 / 5.26	1 / 2.22	10 / 4.06
Cartoon Hero	9 / 11.11	13/29.54	4 /5.26	2 /4.44	28 / 11.38
Others	7/8.64	-	4/ 5.26	3 / 6.66	14 / 5.70
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

In Table 2; 78.86% of the students perceive the teacher as "human." However, about 11.38% of the students perceive the teacher as a "cartoon hero." It is also seen that 4.06% of the students perceive the teacher as a "recognized person" (eg Albert Einstein, M. Kemal Atatürk). As seen in Table 2, it can be said that the students made more realistic pictures. In this case, the fact that a great majority of teachers are portrayed as human beings can be considered as a reflection of reality in the picture. The 11% student group, which is the second highest rate in Table 2, likened teachers more to cartoon characters. This can be explained by the creativity of the students in drawing pictures.

Table 3. Perceptual Gender Perceptions of Students

Perceptual Gender	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
Woman	59/ 72.83	9/20.45	37/48.68	24/ 53.33	129/52.43
Man	17/20.98	27/61.36	36/47.36	17/37.77	97/39.43
Not Human	2/2.46	4/9.09	3/3.94	-	9/3.65
Unknown	2/2.46	3/6.81	-	-	5/2.03
Others	1/1.24	1 / 2.27	-	4/8.89	6/2.43
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

As seen in Table 3, 52% of the students who participated in the survey stated their teachers as women in their paintings. Again referring to Table 3, it is seen that 39.43% of the gender perceptions of teachers are "men" in the pictures drawn by the students. It is seen that about 8% of the students who participated in the research have drawn their teachers in the sub-

materials such as "Not human", "Unknown", "Other" (materials for science lesson instead of teacher). As seen in Table 3, it can be said that the students depicted their teachers as women to a great extent. According to this, it can be deduced that female teachers tend to be more involved in this area in terms of science courses.

Table 4. Physical Appearance of the Teacher

Physical Appearance of the Teacher	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
With Suit	22/27.16	6/13.63	11/14.47	4/8.88	43/17.47
White Apron	11/13.50	11/25.00	29/38.15	17/37.77	68/27.64
With tie	18/22.22	8/18.20	13/17.10	6/13.33	45/18.29
Sweatpants	2/2.46	-	1/1.31	-	3/1.21
Messy	3/3.70	-	2/2.63	2/4.44	7/2.84
Stylish Dress	17/20.98	10/22.72	9/11.84	9/20.00	45/18.29
Young	2/2.46	7/15.90	11/14.47	7/15.55	27/10.97
Not Drawn	4/4.93	-	-	-	4/1.62
Others	2/2.46	2/4.54	-	-	4/1.62
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

As seen in Table 4, about 28% of the students who participated in the research described their teachers as wearing "white overalls" in the drawings they had drawn. Approximately 19% of the students described their teachers as wearing "a tie" and about 19% as "elegantly dressed." In Table 4, it is seen that the student group that depicts the teachers as wearing "white overalls" is the 7th grade students. Beside these, the level of describing teachers as wearing white overalls is progressing in line with the grade level. From here it is also possible to reach the conclusion that the teachers wearing white overalls when entering the classroom increases as the grade level increases.

Table 5. Students' Metaphores for Teachers

Metaphores	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
Sun	2/2.46	-	-	1 / 2.22	3 / 1.21
Book	1 / 1.23	4 / 9.09	1 / 1.31	1 / 2.22	7 / 2.84
Heart	1 / 1.23	2/ 4.54	-	2 / 4.44	5 / 2.03
Moon	-	-	-	-	-
Star	-	1 / 2.72	-	-	1 / 0.40
Cloud	-	-	-	-	-
School	1 / 1. 23	1 / 2.72	-	-	2 / 0.81
Earth	-	-	2/2.63	-	2 / 0.81
Not Drawn	76/ 93.82	36/ 81.82	73/96.05	41/91.11	226/ 91.86
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

From Table 5 it can be seen that most of the students depicted teachers as "books" in their paintings. From the results obtained, it can be seen that the students see their teachers as a source of information like books. It is seen that the students who use metaphors for teachers in their paintings are mostly lower grade students. It can be said that students from the upper grades use more realistic items in their paintings.

Table 6. Teachers' Gestures and Facial Expressions According to Student Perception

Gestures and Facial Expressions	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
Smiling	61/ 75.30	19 / 43.18	36 / 47.36	16 / 35.56	132 / 53.65
Confused	2 / 2.46	4 / 9.09	1 / 1.31	6 / 13.34	13 / 5.28
Excited	3 / 3.70	2/ 4.54	14 / 18.42	2 / 4.45	21 / 8.53
Sad	1 / 1.23	6 / 13.63	7 / 9.21	1 / 2.23	15 / 6.09
Angry	-	1 / 2.72	2 / 2.63	2 / 4.45	5 / 2.03
Shy	-	1 / 2.72	8 / 10.52	-	9 / 3.65
Worried	3 / 3.70	1 / 2.72	-	1 / 2.23	5 / 2.03
Unhappy	4 / 4.93	1 / 2.72	2/2.63	2 / 4.45	9 / 3.65
Considerate	7 / 8.64	7 / 15.90	-	10 / 22.23	24 / 9.75
Not Drawn	-	2 / 4.54	6 / 7.84	5 / 11.12	13 / 5.28
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

Table 6 shows that findings related to the gestures and facial expressions of teachers are seen according to the perceptions of the students. According to this, it can be said that the students perceive the teachers as mostly "happy-faced". From here it can be reached that the teachers have a positive influence on the students during the learning-teaching process.

Table 7. Dimensions of Teacher Figure by Perceptions of Students

Dimenssions	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
Large	5/6.17	6/13.63	2/2.63	3/6.66	16/6.50
Small	7/8.64	9/ 20.45	1 / 1.31	1 / 2.22	18/ 7.31
Realistic	64/79.012	22%50.00	70/92.10	37/82.22	223/90.65
Not Drawn	5/6.17	8/18.18	3/3.94	4/8.89	20/8.13
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

In the pictures drawn by the students seen in Table 7, the size of the teacher figure is realistic by 90%. According to this, it can be said that in the pictures of the students close to the whole, the teachers and the other objects are conveyed on paper with their actual dimensions. Looking at the other subcategories, 7% of the students can achieve the result that they are small with the teacher figure.

Table 8. Physical Characteristics of Teachers from Perceptions of Students

Physical Charasteristics	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
With Eyeglasses	1 / 1.23	4 / 9.09	8 / 10.52	1 / 2.22	14/5.69
Messy Hair	13 / 16.04	9 / 20.45	22 / 28.94	17 / 37.77	61/24.79
Groomed	48 / 59.25	19 / 43.18	38 / 50.00	26 / 57.77	131/53.25
Bald	2 / 46	1 / 2.72	2 / 2.63	-	5 / 2.03
Bearded	-	1 / 2.72	-	-	1/ 0.40
Not Drawn	7/ 8.64	9 / 20.45	4 / 5.26	1 / 2.22	21 / 8.53
Others	10 / 12.34	1 / 2.72	2 / 2.63	-	13 / 5.28
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

In Table 8, perceptions of the students about the physical appearance of the teacher are seen. More than half of the students have shown their teachers "clean and well-maintained". Some students painted their teachers as "hair scattered". Together with these, students did not depict their teachers as having "remarkable injuries" or "physical disabilities." From here it can be said that the students perceive the physical appearance of the teachers as more positive.

Table 9. Location of Teachers by Perceptions of Students

Location of Teachers	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
Classroom	37 / 45.67	14 / 31.81	12 / 15.78	9 / 20.00	72 / 29.26
In front of the Board	27 / 33.33	7 / 8.64	9 / 11.84	3 / 6.66	46 / 18.69
Table	3 / 3.70	2 / 4.54	6 / 7.89	1 / 2.22	12 / 4.78
Near the Board	9 / 11.12	4 / 9.09	14 / 18.42	11 / 24.44	38 / 15.44
In Laboratory	3 / 3.70	11 / 25.00	21 / 27.63	14 / 31.11	49 / 19.91
In a field	-	-	2 / 2.63	1 / 2.22	3 / 1.21
School Garden	-	-	1 / 1.31	3 / 6.66	4 / 1.62
Ceremony	-	-	3 / 3.94	-	3 / 1.21
Activity	-	1 / 2.27	2 / 2.63	3 / 6.66	6 / 2.43
On Computer	1 / 1.23	3 / 6.81	5 / 6.57	-	9 / 3.65
In Front of the Flagpole	-	-	1 / 1.31	-	1 / 0.40
In the Sky	1 / 1.23	2 / 4.54	-	-	3 / 1.21
Not Drawn	-	-	-	-	-
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

Taken into account the Table 9, it is seen that students depicted their teachers more "in class". Approximately 20% of the students depict their teachers in the "Laboratory", while some students depict their teachers "in front of the Board". Again, 15% of the students have shown their teachers "in the picture". From the obtained findings, it can be reached that the teachers continue the learning-teaching process in the class environment and the students also perceive the teachers in this way.

Looking at Table 10, it can be seen how the students conveyed the actions of the teachers according to the perception of the students. As seen in Table 10, about 45% of the students depicted their teachers as "writing on the board," "walking around the school" and "teaching." Despite this, the proportion of students drawing "when doing experiment", "observing", and "when performing activities with students" was found to be very low. The fact that observations and experiments constitute the basic structure of the science course are made so low according to the perception of the students plays a big role in the importance of the research.

As can be seen in Table 11, there are objects in the hands of the teachers in the students' depictions. Teachers who need to have experimental equipment in the laboratory environment and mostly in the science class have "books" in their hands with a rate of 26.82% according to the perception of the students. Approximately 25% of the pupils depicted in their teachers' materials such as "Ruler-stick" and "Pencil". Approximately 13% of the pupils depicted their teachers in their hands with "Student's Hand" and "Flower."

Table 10. Types of Teachers' Actions Perceived by Students

Types of Teachers' Actions	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
Writing on the Board	23 / 28.39	8 / 18.19	11 / 14.47	6 / 13.33	48 / 19.51
Walking in Classroom	14 / 17.89	5 / 11.36	13 / 17.10	4 / 8.89	36 / 14.63
Speaking to Students	7 / 8.61	3 / 6.81	3 / 3.94	1 / 2.22	14 / 5.69
Reading Paper	1 / 1.23	-	-	-	1 / 0.40
Reading Book	-	2 / 4.54	1 / 1.31	3 / 6.67	6 / 2.43
Lecturing	16 / 19.75	9 / 20.45	16 / 21.05	9 / 20.00	50 / 20.32
Experimenting	9 / 11.11	6 / 13.63	12 / 15.78	11 / 24.45	38 / 15.44
Observing	-	3 / 6.81	2 / 2.63	1 % 2.22	6 / 2.43
Showing affection to Students	-	-	1 / 1.31	1 / 2.22	2 / 0.81
Giving Students a Flower	-	1 / 2.27	2 / 2.63	-	3 / 1.21
Playing with Students	-	1 / 2.27	1 / 1.31	2 / 4.45	4 / 1.62
Activity with Students	5 / 6.17	2 / 4.54	2 / 2.63	3 / 3.67	12 / 4.87
While Standing	6 / 7.40	3 / 6.81	9 / 11.84	4 / 8.89	22 / 8.94
Not Drawn	-	-	-	-	-
Other	-	1 / 2.27	2 / 2.63	-	3 / 1.21
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

Table 11. Objects in Teachers' Hands According to Perceptions of Students

Objects in Teacher's Hands	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
Ruler	17 / 20.98	4 / 9.09	8 / 10.52	3 / 6.67	32 / 13.00
Book	21 / 25.92	15 / 34.09	19 / 25.00	11 / 24.44	66 / 26.82
Bag	6 / 7.40	2 / 4.54	6 / 7.89	-	14 / 5.69
Paper	12 / 14.81	-	5 / 6.65	3 / 6.67	20 / 8.13
Chalk	-	-	1 / 1.31	-	1 / 0.40
Flower	7 / 8.64	6 / 13.63	5 / 6.57	1 / 2.27	19 / 7.72
Pencil	2 / 2.46	7 / 15.90	11 / 14.47	10 / 22.72	30 / 12.19
Ball	-	-	-	-	-
Rod	2 / 2.46	3 / 6.81	7 / 9.21	4 / 8.89	16 / 6.50
Palette	-	-	-	-	-
Student's Hand	4 / 4.93	1 / 2.72	3 / 3.94	2 / 4.45	10 / 4.06
Not Drawn	7 / 8.64	6 / 13.63	11 / 14.47	9 / 20.00	33 / 13.41
Others	3 / 3.70	-	-	2 / 4.45	5 / 2.03
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

Table 12. Objects Surrounding the Teachers by Pupils' Perceptions

Objects Surrounding the Teachers	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
Flag	8 / 9.87	12 / 27.27	6 / 7.89	3 / 6.67	29 / 11.78
School	2 / 2.46	1 / 2.27	5 / 6.57	1 / 2.22	9 / 3.65
Students	11 / 13.58	2 / 4.54	6 / 7.89	5 / 11.11	24 / 9.75
School Garden	2 / 2.46	-	-	1 / 2.22	3 / 1.21
Book Shelf	14 / 17.28	3 / 6.81	8 / 10.52	7 / 15.56	32 / 13.00
Board	17 / 20.98	11 / 25.00	19 % 25.00	10 / 22.23	57 / 23.17
Table	9 / 11.11	2 / 4.54	11 / 14.47	7 / 15.56	29 / 11.78
Atatürk's Corner	5 / 6.17	5 / 11.36	7 / 9.21	6 / 13.34	23 / 9.34
Flowers	1 / 1.23	2 / 4.54	-	-	3 / 1.21
Star	2 / 2.46	-	1 / 1.31	-	3 / 1.21
Test Tubes	7 / 8.64	6 / 13.63	12 / 15.78	4 / 8.89	29 / 11.78
Others	3 / 3.70	-	1 / 1.31	-	4 / 1.62
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

From Table 12, when looked at the perceptions of the students about the objects that are around the teachers, it is seen that 23.17% of the students depict "Board" around their teachers. This is followed by "Book Shelf" with 13.00% and "Flag" with 11.78%. The "test tubes", which are the first materials that should come to mind about the science course, are among the objects drawn around the teachers with a ratio of 11.78%. From this data, it can be suggested that teachers hold lessons in the classroom environment rather than in the laboratory environment in the process of teaching and learning science lessons.

Table 13. *Seating Patterns According to Students' Pictures*

Setting Patterns	Fifth Grade f / %	Sixth Grade f / %	Seventh Grade f / %	Eighth Grade f / %	Total f / %
Traditional Rows	26 / 32.09	17 / 38.63	22 / 28.94	19 / 42.22	84 / 34.14
Semi Circle	9 / 11.11	5 / 11.36	7 / 9.21	2 / 4.44	23 / 9.34
U Scheme	32 / 39.50	9 / 20.45	14 / 18.42	7 / 15.56	62 / 25.20
Cluster Configuration	3 / 3.70	1 / 2.72	4 / 5.26	1 / 2.22	9 / 3.65
Free	8 / 9.87	7 / 15.90	17 / 22.36	6 / 13.34	38 / 15.44
Ceremony	2 / 2.46	1 / 2.72	2 / 2.63	5 / 11.12	10 / 4.06
Not Drawn	1 / 1.23	3 / 6.81	6 / 7.89	3 / 6.67	13 / 5.28
Others	-	1 / 2.72	4 / 5.26	2 /4.44	7 / 2.84
Total	81/32.92	44/17.89	76/30.89	45/18.29	246/100.0

The findings given in Table 13 reveal that the perceptions of the layout of the students are drawn by the students. According to this, 34% of the students who participated in the survey depicted their seating styles as "traditional rows" order. 25% of the students depicted their seating layout as "U-shape", but this ratio is quite low. Again, as many as 15% of the students have illustrated their seating layout as "free". The fact that the ratio of the free seating order is so high can bring criticism to mind either positively or negatively. Here, the communication between the teacher and the students is an important point where they prefer free seating because of the intention to increase inter-class interaction or lack of competence in class management.

Table 14. Objects in Classroom Based on Perceptions of Students

Objects in Classroom	Available f / %	Not Available f / %
Board	194 / 78.86	52 / 21.13
Table	202 / 82.11	44 / 17.88
Desks	185 / 75.20	61 / 24.79
Ataturk Portrais and National Anthem	163 / 66.26	83 / 33.73
Panels	177 / 71.95	69 / 28.04
Projector	38 / 15.44	208 / 85.55
Overhead	7 / 2.84	239 / 97.15
Computer	32 / 13.00	214 / 86.99
Test Materials	24 / 9.75	222 / 90.24
Models	13 / 5.28	233 / 94.71
Flag	169 / 68.69	77 / 31.30

Table 14 shows that there are objects in the class according to the perceptions of the students. A large majority of students depict the classrooms with objects such as "board", "table", "desks", "pin boards," reflecting the traditional classroom environment. A large majority of the same students did not show the pictures of "Computer", "Projection", "Overhead" and "Experimental Materials" in their drawings. Their pictures, which constitute a more technological classroom environment, support more permanent learning and teaching environment. It can be said that the teachers who teach the science course are not using the class environment effectively and cannot integrate the technology into the classroom environment.

4. DISCUSSION and CONCLUSION

As a result of the research carried out, it has been ascertained that the students perceive teachers as "human beings" to a great extent, and they portray them as such. However, some students perceive their teachers as "cartoons" and others as "well-known people." A group of students used metaphors while drawing their teachers, likening their teachers to the sun or the stars. From here it can be said that a large majority of the students are realistic in perceiving their teachers.

When the metaphors used by the students are examined in detail, it is seen that the metaphors used have an important place in human life. The fact that students transfer their teachers as important assets in this way shows that they have positive views of the teachers. However, it can be concluded that they perceive their teachers as a source of information.

When the students perceptions of the teachers' gender are examined, it is understood that the figures which are depicted with a small proportion are mostly female teacher figures. From this point of view, it is possible to reach the conclusion that the students have more courses in science lessons with female teachers and at the same time, female teachers prefer to teach more in sciences than male teachers. In view of the data obtained and examined in the survey, it is seen that the students mostly depicted their teachers as wearing "white overalls" when they perceived the teachers' physical appearance. Some students portrayed their teachers in "suits" and "ties" and as "stylishly dressed." Accordingly, it can be said that the students did not show the teachers more as white doves, so that the teachers were able to reflect more in the laboratory environment, or at least to reflect the science teachers' view of their students. In the study of Aykaç (2012), it has been seen that the students in the same subcategory draw their teachers more in "suits" and as "elegantly dressed." In both surveys, the physical appearance of the teachers can be interpreted in such a way that the teachers have a positive effect on the students.

According to the perceptions of the pupils, when looking at the dimensions of the teacher figure, it is seen that the students are mostly "realistic" when drawing their teachers. When the age group of the students participating in the research is taken into account (10-15), the teachers are closer to realistic dimensions in the drawings of the students. In Aykaç's work, it is seen that the students draw pictures with more realistic dimensions.

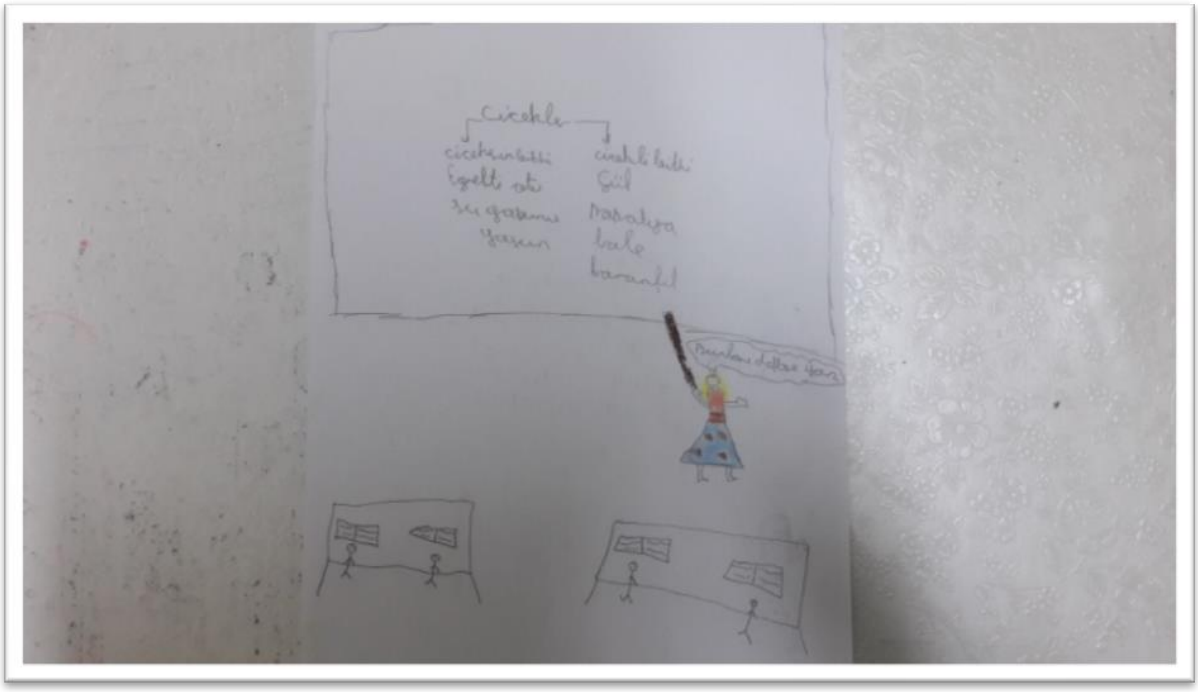
In addition to these, some students in the Aykaç study have been able to see that while the teacher has been drawn larger and smaller than realistically, the students are more inclined to draw their teachers as smaller rather than larger.

Picture 1. *Drawing of 108 Coders from 7th Grade*

As you can see in Picture 1, students are more likely to make small presentations than to draw the teacher large. As a result, it can be deduced that teachers are inadequate in the classroom or laboratory environment, failing to address all students, manage the classroom, and impliment the learning-teaching process. When the findings of the teachers' gestures and facial expressions were examined, it was seen that the students portrayed their teachers as happy faced. From here it is possible to reach the conclusion that teachers have a positive effect on students.

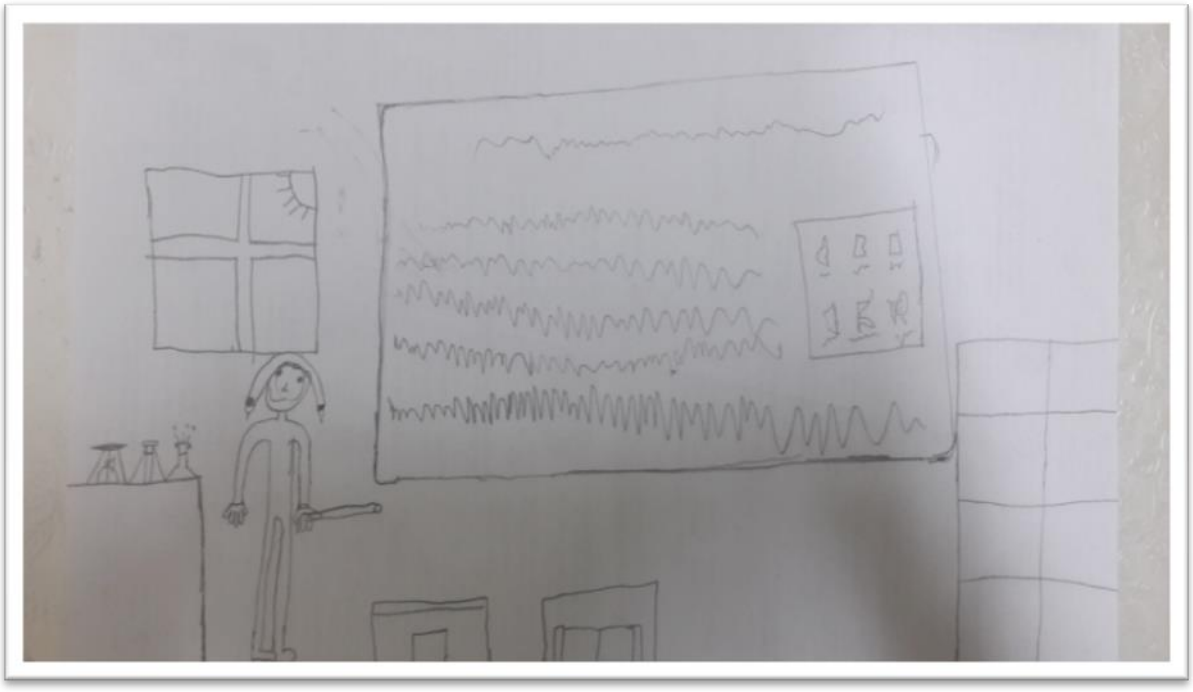
When the students' peceptions of the teacher's physical characteristics were examined, it was found that the students described the teachers as "clean and well-maintained." From this, it can be said that teachers have positively affected the students in terms of physical appearance. But, some students portrayed their teachers with "scattered hair" and it is inevitable that some teachers are a negative example in terms of physical appearance. According to the research, one of the most important findings is that the places where the teachers are located are more in class and in front of the board.

Looking at Picture 2, one can see that the science teacher is depicted in a traditional way, that is, in front of a book, in a classroom arranged by traditional order, while it should have been in a way that a science teacher should be perceived more in a laboratory environment or in places such as gardens, museums, or science-art centers.



Picture 2. *Drawing of the Learners Coded as 64 in 5th Grade*

In science education, the teacher should know how to create learning opportunities with organized activities both inside and outside the classroom, and to extend the learning-teaching process so that every student has opportunities created for them (Ayvacı & Ünal, 2017). The fact that science teachers are depicted in the traditional classroom environment even though they should have been portrayed more likely in a laboratory or outdoors shows that they cannot expand their role in the learning-teaching process and cannot use the lab environment effectively in science teaching. The representation of teachers in the highest grade as "in front of the board" is also an indication that teachers cannot manage the learning-teaching process, or take into account the students' individual differences, and try different teaching methods.



Picture 3. *Drawing of 88 Coded student from 6th grade*

When teachers' actions are considered, the teachers are depicted more as standing during lessons as shown in Picture 3. If this situation is to be evaluated in terms of science, it can be said that science teachers do not perform experiment activities and activities that support students' learning by doing the most important thing in science and strengthening relations among students and taking into consideration the classroom or laboratory environment they are in. The objects in the teachers' hands also provide us with important clues as to how they direct the learning-teaching process. According to the research, mostly books were displayed in the teachers' hands. From here, it is possible to say that teachers mostly benefited from the books as resources in the class environment. Today, with the development of technology, the learning-teaching process and the education-learning environments with it also change. It is expected that teachers will benefit from the most technological advancements in the learning-teaching process and to make the technological tools and equipment a continuous part of the classroom environment in an effective way. According to the research findings, teachers do not include these tools in the learning-teaching process, and still perform teaching activities by traditional methods.

In the learning-teaching process, tools are generally used to support teaching. Well designed and useful materials enrich the teaching environment and increase the quality of teaching together with it.

Picture 4. *Illustration of 224 Coded Learners from 8th grade*

Tools used in the process provide a multi-learning environment and contribute to meeting the individual needs of the students. Tools are used to attract attention, facilitate remembrance, embody abstract learning, enable time saving, re-use, and increase understanding by simplifying content (Yalin, 2012; 82-90). When the objects in the classroom environment shown in Picture 4 are examined carefully, it is seen that most of the students draw materials that can be found in almost every classroom while it should have depicted enriched teaching environments and shown materials to meet individual needs. From this, it is concluded that materials which enrich the course content and help simplify the process are not used enough. It is necessary to use these materials more effectively in the classroom and laboratory environment during the learning-teaching process.

Also, the research reveals that the enrichment of teaching and individual differences is affected by the seating layout of the students in the learning-teaching environment. Regulation of the classroom environment increases the quality of teaching and helps students to learn easily (Yalın, 2012; 103). If the findings of the classroom are interpreted according to the perceptions of the students participating in the research, it can be said that the classic seating is mostly used in the classroom. Communication in the classroom is the lowest level in the traditional seating plan. However, it is not possible to use discussion techniques effectively in this order (Yalın, 2012; 103). Not only for the science class, but also for the classroom or laboratory environment, the most recommended is the U-shaped seating arrangement. Classroom interaction increases in the U-class seating arrangement, which enables students to have better quality interactions with each other. A teacher's preference for traditional seating order may indicate the inadequacy of teachers' knowledge of classroom management and classroom organization, or that their classes are too crowded to implement it. The arrangement of the teaching environment should not be limited to the traditional seating arrangement only.

Picture 5. Illustration of 184 Coded Learners from 8th grade

The use of equipment in the teaching environment and in the learning-teaching process is also very important. In the course of the research, the objects in the classroom have also been studied. Students also depict objects such as Ataturk Portraits, National Anthem, Turkish Flag, which are traditionally found in Turkish classrooms, as well as objects such as projectors, computers and overhead projectors as shown in Picture 5. Unfortunately, the number of students painting these tools remains very low. From this, it can be acertained that there is need for essential tools in the learning-teaching environment, but they are not used effectively.

This analysis of pupils' perceptions of their images indicates that the physical appearance of teachers in general has a positive effect on students overall and that technological tools and that equipment is not used well in the classroom environment. But, technology can adapt to teaching environments very quickly. However, in science class, it has been found that the learning-teaching process is still mostly done in the classroom environment, and that students can participate very little in classroom communication by sitting in the traditional seating order. From this point of view, it has been revealed that in the science classes, teachers are still lacking

in the learning-teaching process and have problems in the effective use of classroom management, teaching techniques and materials.

In science classes, teachers need to integrate information technologies well into the learning-teaching process in order to produce more qualified learning-teaching processes. So as to provide more qualified and lasting learning, teachers can better analyze the emerging technology and integrate it well into the learning-teaching process. In addition, the seating layout of the classroom is also very important in planning the learning-teaching process. Teachers should opt for a U-shaped seating arrangement in the classroom to enhances and facilitate teacher-student, student-student communications.

In today's world where the technological developments and knowledge change rapidly, the seating order in the classroom, the processing of science lessons in the traditional classroom environments becomes meaningless. Instead, teachers should choose to conduct science lessons in places that will create richer learning opportunities, such as laboratories, museums, science-art centers, school and outdoors rather than conducting science classes in a traditional classroom settings. It will be more useful to evaluate the results of this study not only within the content of this research work, but also within the scope of all the courses carried out throughout the country in order for the individuals trained to adapt to the developing world.

5. REFERENCES

Artut, K. (2004). [An investigatıon on children's development level of drawing during pre-school illustration education]. *Journal of Çukurova University Institute of Social Sciences*, 13. (1). 223-234.

Aykaç, N. (2012). [Perceptions of the Teacher and Teaching Process in the Drawings of Elementary School Student]. *Education and Science*, 37; 364, 298-315.

Ayvacı, H.Ş. & Ünal, S. (2017). *Kuramdan Uygulamaya Okul Öncesinde Fen Eğitimi*, Pegem Akademi, Ankara.

Burden, P.R. (1995). *Powerful Classroom Manageme].t Strategies: Motivatring Students to Learn*. Corwinn Press. California.

Büyüköztürk, Ş, Çakmak, E,K; Akgün, Ö,E, Karadeniz, Ş, Demirel, F (2009). *Bilimsel Araştırma Yöntemleri*. Pegem Akademi, Ankara.

Clark, H. (2002) *Building education: The role of the physical environment in enhancing teaching and research*, London: Institute of Education, University of London

Cohen, L., Manion, L., Morrison, K. (2000). *Reseach Methods in Education*. 5th Edition. Routledge Falmer.

Corbett, D., &Wilson, B. (2002). What urban students say about good teaching. *Educational Leadership*, 60(1),18-22.

Dudek, M. 2000. *Architecture of schools: The new learning environments*, London: Architectural Press.

Gökçe, F. (2014). *Sınıfta Öğrenme ve Öğretme Sürecinin Yönetimi*. Pegem Akademi. Ankara.

Gültekin, Ç., Tosun, Ö., Turgut, Ş., Örenler, Ş., Şengül, K. and Top, G. (2010). Promoting an inclusive image of scientists among students: Towards research evidence-based practice. *International Journal of Science and Mathematics Education National Science Council*, Online. Taiwan.

Hsiao, C.Y. & Chen, C,M. (2015). Examining Kindergarteners' Drawings for Their Perspectives on Picture Books' Themes and Characters. Canadian Center of Science and Education. *International Education Studies*; Vol. 8, No. 11; 2015 ISSN 1913-9020 E-ISSN 1913-9039.

Jacobsen, D. A., Eggen, P., Kauchak, D. (2006). *Methods for Teaching. Promoting Student Learning in K-12 Classrooms*. 7th edition. PearsonEducation. New Jersey.

Johnson, B., Christiensen, L. (2012). *Educational Research Quantitaive, Qualitative, and Mixed Approaches*. Fourth Edition. Sage Publication.

Lin, M. Y. (2006). *Appreciation And Application Of Picturebooks*. Taipei: Psychological Press.

Lowenfeld, V., & Brittain, W. (1987). *Creative And Mental Growth* (8th ed.). New York: Macmillan.

Malchiodi, C.A. (1998). *Understanding Childrens' Drawings*. GuilfordPress. New York.

Moore, K. D. (2001). *Classroom Teaching Skills.* McGrawHill. 5th edition. New York.

Skoumios, M., Kambouropoulou, M. S. (2012). Investigating Pupils' Images of Science Teaching Using Drawings. *The International Journal of Science in Society* Volume3,Issue2,2012,http://science-society.com/journal/,ISSN1836-6236.

Thompson, G. (2002). African American teens discuss their elemantry teachers. *Educational Horizons*, 80(3), 147-152.

Weber, S. J., & Mitchell, C. (1996). Drawing ourselves into teaching: Studying the images that shape and distort teacher education. *Teaching and Teacher Education*, 12(3), 303-313.

White, R., & Gunstone, R. (1992). *Understanding Probing*. Routledge. London.

Yalın, H.İ. (2012). *Öğretim Teknolojileri ve Materyal Geliştirme,* Nobel Akademi, Ankara.

Yıldırım, A; Şimşek, H (2000). *Sosyal Bilimlerde Nitel Araştırma Yöntemleri*. Seçkin. Ankara.

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Evaluating the Comparability of PPT and CBT by Implementing the Compulsory Islamic Culture Course Test in Jordan University

Abdelnaser Sanad Alakyleh

To cite this article: Alakyleh, A. S. (2018). Evaluating the Comparability of (PPT) (CBT) by Implementing the Compulsory Islamic Culture Course Test in the University of Jordan. *International Journal of Assessment Tools in Education*, 5(1), 176-186. DOI: 10.21449/ijate.370494

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 176-186

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Evaluating the Comparability of PPT and CBT by Implementing the Compulsory Islamic Culture Course Test in Jordan University

Abdelnaser Sanad Alakyleh

Ministry of Education in Jordan/ Formerly Al-Jouf University

Abstract: Study aims to determine whether the university students' scores in the compulsory Islamic culture course test on a selected sample differ across the paper-and pencil test (PPT) & computer-based test (CBT) versions, and to reveal the relationship between gender and the student's level of performance in the test, Therefore, the study evaluated the comparability of two versions of a compulsory Islamic culture course test (PPTs) and (CBTs). The importance of conducting the study in Jordan stems from the fact that public and private universities have begun to move away from the traditional patterns of tests such (PPTs) and went towards (CBTs), In addition to detecting which models give the best in the output and has the characteristics of the psychometric test, Furthermore indicates whether there were differences between males and females, the study sample consisted of 120 individuals, 67 females and 53 males from scientific, health and humanities colleges. The results showed that there was no significant difference between the two versions provided to students CBT and PPT with 0.36 moderate correlation indicators in the pre-CBT test, no significant differences between the males and females in the CBT test results. Therefore, on the basis of the results of the present study, the CBT test is an option and a preferred alternative for regular students of the bachelor's level at the University of Jordan.

ARTICLE HISTORY
Received: 15 September 2017
Revised: 16 December 2017
Accepted: 20 December 2017

KEYWORDS
PPT,
CBT,
Comparability,
Gender difference,
Test preference

1. INTRODUCTION

CBT has recently appeared as one of the most demanded viable form of alternative assessment throughout the world. Along with the development of computer assisted language learning (CALL) in education, applying computers as accepted assessment tools seem to be inevitable especially in academic settings. In education, CBT is used to evaluate the language proficiency of English learners (Fleming & Hiple, 2004). Also, computer-based testing CBT has grown in popularity and will likely become the primary mode for delivering tests in the future. Computers revolutionized the world of training and development. Many investigators such as Fuhrer (1973) began researching on many point of mode which has enhanced training through computers. Many studies focused on the effects of using computers in the classroom for testing on various aspects of the learning environment such as student anxiety, teacher attitudes, student achievement and more.

*Corresponding Author E-mail: abd_275@yahoo.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.370494**

The computer had a significant impact on education over the past 20 years, its impact on educational testing is interesting and remarkable, although a number of large educational institutions, such as ETS and English, Cambridge ESOL has designed CBTs, a limited number of educational institutions have adopted these tests, and few teachers apply them to their students, which explains the continued dominance of PPTs on the educational field. In the current period, the development of science and technology is advancing. This has an impact on life, including on education. The presence of technology in education is used to assist and improve the quality of learning (Woolfolk, 2007)., while the number of countries regard education as crucial for improving their current situation in every respect and moving it a step further in the information age of the 21st Century. In this context, Aslan (2006) pointed out that the developments which have occurred in information technology have given students fast and easy access to information, which has made a great contribution to education systems. As an example of the accelerated use of computers in the educational and academic fields mostly in tests, there have been several different versions of these examples and applications that have become issues of interest to researchers and those interested in the field of educational and academic applications in the field of tests and comparisons with methods and traditional versions used by educators and academics to submit to examiners. With a view to carrying out the assessments of the examiners through its results on the applicable tests version in an effort to improve the quality and accuracy of subsequent decisions.

It is important to address two types computer-based tests, Computer based standard testing CBTs and Computer-adaptive testing CATs, the CBT test is, in short, the usual paper version of the test, which has been converted into CBTs. Therefore CBT is as static as in the original paper copies of the test. In other words, all applicants for the computer test answer the questions in the same order in which questions are presented in the paper version, while a computer test adapted to the language proficiency of the CAT student, the tester’s answer different sets of questions, which are asked according to their level. Their answer affects a question about the following questions. A little bit of the first, and put it on the applicant to the test, and vice versa if the answer is wrong, the computer will choose an easier difficult question, hence the name "adaptive test". CBTs are characterized by a number of features, tests are more stable and credible, and the CBT is superior to paper testing in many positive aspects. CAT has the ability to perform more rigorous and credible tests in determining the level of language knowledge among students. This is because it uses statistical analysis to assist the language test in identifying weak and good questions (Niemeyer, 1999)., but the problem with computerized tests arises when the matter of validity comes; however, there is no evidence to show that the construct of CBT may produce less valid tests. Instead, other factors may influence tests that have little to do with the testing objectives the test developer intends to provide. For example, in many CBTs, it seems that the test designer started from a valid objective, but the limitations of the program, system, language or the tester's own characteristics have influenced the results of tests (Chapelle & Douglas, 2006).

(Khoshsima1 et al., 2017) He explained that CBT has recently appeared as one of the most demanded viable form of alternative assessment throughout the world. Along with the development of computer assisted language learning (CALL) in education, applying computers as accepted assessment tools seem to be inevitable especially in academic settings, as mentioned (Holtzman, 1970) that IBM version 805 machine used in 1935 has been recorded as the first attempt to use computers in educational testing domain. It aimed to score objective multiple-choice item tests of American test takers each year to reduce the costs of scoring labor of millions of test takers throughout the USA, after publication of the first book on CBT in language domain, also (Al-amri, 2009) pointed out that many developments in technology caused rapid enhancements in comprehensive language testing software packages to use great advantages of CBT such as the innovation, efficiency and productivity, CBT assesses test

taker's language proficiency accurately by providing more efficient standardization of test administration conditions, in CBT the same instructions, materials and information are presented in an enhanced consistent and uniform way to all test takers, regardless of the testing population size, place and time of testing. Moreover, unlike paper examination in conventional classroom, immediate viewing of scores on screen is provided in CBT session to give test takers the instant feedback. But, in some cases of large-scale CBT occasions, the security issues such as identity detection of test takers are the main concern.

Universities, some institutions, and testing organizations have started to change the mode of testing administration and to replace their paper and pencil tests PPTs with CBTs in language assessment field (Kate, 2012), while comparability and equivalency of test scores between the two test administration modes have been the real concerns for educators, scholars, practitioners and designers in assessment field (Lottridge, Schulz, 2008).

The sequence of studies and research on the preference of the examiners and educators to PPT compared to CBT, such as (Ackerman, 2011; Clariana, 2005; Creed et al., 1987; Destefano, 2007; Dillon, 1994; Dundar, 2012 & Monirosadatet et al., 2014) study, showed that they agreed to prefer computer examiners CBT, while their results are better on paper and pen PPT, while (Higgins et al., 2005), (Al-amri, 2009) have been mentioned that there are no significant differences between the use of both models nor correlation between test mode preference and testing performance, used in the test and the performance of students, with regard to the gender of the respondent and his preference for any of the two models, some studies, such as (Gallagher, et al., 2002; Wallace & Clariana, 2005) indicated a preference for females to use the form PPT in front CBT model.

From the review of educational literature and previous studies that dealt with this important issue, the results of no significant differences between the use of both models, used in the test and the performance of students, shortage of correlation between test mode preference and testing performance, remains a subject of discussion and extensive examination of the different variables that affect the results of both versions such as gender, ethnic variables, motivation of the examiner, the concern of the test, the conditions of application, cognitive processes and technical issues which lead to the conclusion and the result that the use of the computer is not the tool of choice for evaluation, computers have become more widespread and used in academic aspects, especially in the application of tests in all its forms and their versions and in the results of which they depend on mainly the analysis of important decisions academically and practically, it has produced a lot of studies in the field of comparison between CBTs and paper and PPTs results are not compatible or consistent in the field of validity, reliability and significance differences of test scores.

Therefore, based on the above, the current study was to follow up and complete the research and study carried out by the researchers on the use of test models based on PPT compared to using CBT applying both models to a sample of university students and to a completely different topic of language, which focused on most studies in the application, and based on the availability of data and the potential and desire of volunteers from university students, the study came to discuss the comparison between the models of application on the subjects to confirm or deny or modify the previous studies of the results and analyzes.

2. METHODOLOGY

2.1. Purpose of The study

What may affect the validity of the effects of the test mode and the reliability of those results are not specific since the subjects of both male and female gender and their preference to test mode and performance will continue to discuss and research that the results of studies have varied between agreement and conflict on the subject and perhaps the proliferation of

computers in individual everyday life and make life more Automation. The increase in the use of computers in the academic community, especially in the field of tests, requires that traditional tests such as PPTs compared to CBTs waste time in preparation, processing, assessment and effort, as well as the tendency of the subjects often to computer tests. Equating the scores received from two types and suppressing test management, this may require further research on the relationship between some external variables of the mediator such as the sex test and test mode with test performance with greater attention, so the present study aims to determine whether the university students' scores in the compulsory Islamic culture course on a selected sample differ across the versions and to reveal the relationship between gender and the student's level of performance in the test, based on this purpose, the study derived the following questions:

RQ1: Is there any statistically significant difference between PPTs and CBTs when applying of the Islamic culture course test for students of the University of Jordan?

RQ2: Is there any significant difference in test results of CBT between female and male to Islamic culture course test on the students of the University of Jordan?

RQ3: Do performance on CBT affected by participants' prior testing mode preferences?

2.2. Method

The present research that covered both comparison and correlational studies explored the comparability of paper and computer-based testing in a compulsory Islamic culture course and the correlation between some external moderator factors including test taker's characteristics such as computer attitude,. In order to reach solid conclusions in this research, a quantitative instrument's were used to investigate the difference between test results due to its advantages such as easy and fast data collection, consistency and accuracy of collected data and proper descriptive and inferential results, the study used the technique used by the (Khoshsima1, et al., 2017) study to examine the differences between the averages. The analysis of variance ANOVA was used in the study, with the different study population, sample size and nature of the test subject, and to reach the goals of the present study, a quantitative approach including descriptive statistics and was used to answer the first research question by comparing the means of sets of scores and to examine the significant difference between computer familiarity and attitudes, and testing performance of students, add to see if there was any difference between the scores of PPT and CBT. A majority of research conducted on PPT and CBT comparability study focused on the differences in means and standard deviations, (e.g. Makiney, Rosen, & Davis, 2003; Pinsoneault, 1996).

2.3. Population and sample study

The current study society consists of all the students of the University of Jordan for the academic year 2016/2017, which are 35359 students according to the Department of Admission and Registration at the University. The study sample consist of 120 students of both sexes from three faculties chosen by the simple random method with (67 females& 53 males) to ensure that the study community accurately represents the characteristics of the study community as well as and equal opportunities for the appearance of any student from the study community in the sample. The faculties of pharmacy, science and Sharia were selected from the health, scientific and human faculties respectively, according to the conditions of the test and the students' opinions to participate in the experiment until the final stages. As for the reason for selecting the number 120 for the size of the sample, the arithmetic average of one division was taken within the different faculties and there were 40 students. Therefore, for three selected colleges, 120 students were taken, the final number of the study sample. And how to invite these students to participate in the study has been the number of volunteers from colleges, the three who participated in the desire and fill their will and of both genders, male and female,

with the news of the nature of the study and its purpose and mechanism of procedure and applied conditions, the study sample agreed to participate in it.

2.4. Study instrument:

The current study used the final test of the Islamic culture course, which is a compulsory university requirement for all students. To compare the scores from both the CBT and PPT versions, PPT of the Islamic culture course was transferred to the computerized – based version that students will use when they sit for the final test. Another instrument to collect the research data concerning the third research question was a simple question mentioned at the bottom of test takers' exam paper and screen, i.e. would you prefer taking the test on: paper – no difference – computer.

2.5. Procedure:

The method of study begins in the first session of the final test. The students are given the PPT test form using the multiple-choice test format, which includes each item with five options: strongly agree, agree, neutral, disagree and strongly disagree. After the test, the students answered the question: Would you prefer taking the test on: paper-no difference–computer, this question may explore and illustrates the relationship between the preferred version of the test and the performance on the test, while the responses of the students examined were collected and scored. In order to eliminate overwork and stress from the effects of testing and the impact of experience and training and reduce it, the test was done on the computerized – based version after six weeks of testing PPT where the examiners explained oral and written instructions for students to test the computer version. The vast majority of Examine students have demonstrated understanding and prior knowledge with such instructions and how to respond to this type of testing. Each student was given 40 minutes to answer 60 items, with attention to not counting the time of oral and written instruction. The mechanism was to show only one item on the student's test screen. As with the PPT, the examiners have the option to return to any item for review and change the response in the computerized – based version test, the question of the third question was answered exactly as in the first phase of the test at the end of this test.

3. FINDINGS

After the testing and data collection and correction, statistical analysis was carried out using the statistical package for social sciences SPSS V: 22 was the first to verify the validity of the test submitted to the students through the experts validity. The test was presented to a group of specialists in the course content and measurement & evaluation specialists to make their observations on the test items, some of which were deleted or modified while the rest of the test items were kept by the Experts as they are, for the final test to remain in the 60 items. As for the reliability of the test, and because of the importance of the internal consistency of the study data collection instrument, the persistence of a Cronbach's α reliability method was calculated from the test results applied to the Examine students and the test versions, the results of the analysis were shown relatively high reliability coefficients (PPT, α=0.91 and CBT, α=88) (Table 1).

Table 1. Internal consistency reliability (Cronbach's coefficients of PPT & CBT)

Testing Mode	N of Questions	Cronbach's Alpha
PPT	50	0.91
CBT	50	0.88

The sample of the study was divided into 67 females and 53 males. In order to arrive at the answers to the current study questions, the analysis of the ANOVA was used by comparing means of sets of scores to reveal whether there were any differences between the grades of CBT and PPT. Perhaps the most important thing in the current study in the comparison is to find differences in means and standard deviations. With a relatively higher mean score for PPT than for CBT by 0.57 points (Table 2), also (Table 2) shows that the mean scores and standard deviations on the PPT version were (M=53.43, SD= 3.86), while they were relatively lower on the CBT version with (M = 50.12, SD = 3.06). We also note that the standard deviation of the PPT version is higher than that of the CBT version, which means the dispersion of scores from mean score in PPT was higher than in CBT, leading us to conclude that the Standard Error of Measurement (SEM) in the PPT version Above it in the CBT version, This means statistically that a more consistent version in its scores with less dispersion and standard deviation than a PPT version.

Statistical analyzes in (Table 4) showed that there are no significant differences in the scores between the two versions CBT& PPT at the level of statistical significance 0.01. Which supports the null hypothesis that there are no significant differences in the results of the Islamic culture course tests for the two versions CBT& PPT on the students of the University of Jordan.

Table 2. Descriptive Statistics

	N	Mean	S.D	S.E	99% C.I Interval for Mean	
					Lower Bound	Upper Bound
PPT	120	53.43	21.36	3.86	49.65	57.51
CBT	120	50.12	16.74	3.06	48.24	52.00
Total	240	51.78	19.05	3.21	50.76	52.80

The results of ANOVA analysis of the test sessions conducted on the subjects indicated that the significant value was 0.904 at P > 0.01. As this value reveals and illustrates disclosed no statistical significant differences between the scores of test groups resulting from the forms of the test in addition to that the scores of the respondents, also did not differ for the two versions at P < 0.05. Thus, the statistical analysis presented in (Table 2) shows that there are no statistically differences between the PPT version scores of the test (n= 120, M=53.43, SD= 3.86) and the scores of CBT version of the test (n = 120, M = 50.12, SD = 3.06), (Sig = 0.904, p> 0.01).

Table 3. ANOVA Results (Comparison of test scores received from PPT & CBT versions)

	Sum of Square	D.F	Mean Square	F	Sig
Between Groups	5.824	1	5.824	0.013	0.904
Within Groups	16252.667		118	173.734	
Total	16266.154		119		

As for the question of the second study to show whether the scores of the CBT version for the female examiners differ from the results of the degrees of male examiners for the same version, in (Table 4) we note that the distribution of male and female test scores using the CBT version showed that the mean scores of male examiners have reached (M=52.43, SD= 28.36) which is relatively lower than the observed values of females who have reached (M=53.62, SD= 9.74), so the highest mean score was found in Female CBT, with a relatively higher mean score by More than one (1) point slightly. Conversely, the standard deviation of females was lower than that of males from the groups that provided the test CBT, which meant that the test

scores of females were higher than that of males on the CBT version; this raises the values of SEM of female test scores in CBT.

Table 4. Descriptive Statistics (distribution of male and female CBT scores)

	N	Mean	S.D.	S.E.	99% C.I Interval for Mean	
					Lower Bound	Upper Bound
Male CBT	53	52.43	28.36	3.14	42.36	62.50
Female CBT	67	53.62	9.74	2.56	40.55	59.69
Total	120	51.28	26.94	2.23	39.86	62.70

As for the results of the analysis in (Table 5) of the scores of male and female examiners using the CBT version, it shows that the observed significant value was 0.884. This amount of the significant value at 119 (N-1) of degrees of freedom shows no significant differences between the two groups of scores at level 0.01. (Sig= 0.884, p>0.01), thus, one way ANOVA analysis showed that the differences between the male participants' scores in CBT version (n = 53, M = 52.43, SD = 28.36) and female participant scores in CBT version of the test (n = 67, M = 53.62, SD = 9.74) were not statistically significant. (Sig= .884, p>0.01).

Table 5. One-Way ANOVA comparing male and female CBT scores

	Sum of Square	D.F	Mean Square	F	Sig
Between Groups	6.224	1	6.224	0.033	0.884
Withein Groups	6355.224	118	53.86		
Total	6372.194	119			

As for the preference of the test version and the performance of the test and to show the relationship between them, the study examined the Pearson product-moment correlation to reveal this relationship, the results shown in (Table 6) showed that there is moderate correlation of 0.36, which indicated the classification of (Evan, 1996), which means that the changes in pre- CBT preference were Moderately correlated with changes in examine scores on the CBT version. These results differ in terms of the existence of indicators of moderate correlation values with (Flowers et al., 2011, Higgins et al., 2005; & Khoshsima et al., 2017) results for the existence of weak indicators correlation values. This may be due to the difference in the subject of the test in that it has changed from language content to culture content as well as an increase in the sample size used by the current study in which the sample size was 30 individuals, of whom six (6) were female only in (H, Khoshsima et al., 2017) study as an example, but not limited to most of the studies reviewed by the literature of the current study.

Table 6. Pearson product-moment correlation

Pre-CBT testing mode	Pearson product-moment correlation	- 0.36
Preference	Sig (2-tailed)	0.502
	N	120

Correlation of pre-CBT testing mode preference and mean of CBT scores.

The study examined the Pearson product-moment correlation to reveal this relationship between post-CBT testing mode preference and CBT testing performance, the correlation results of the test group in (Table 7) showed no significant correlation, the correlation coefficient of Pearson observed from the analysis was weakly with amount of -0.143.

Table 7. Pearson product-moment correlation

Post-CBT testing mode	Pearson product-moment correlation	- 0.143
Preference	Sig (2- tailed)	0.462
	N	120

Correlation of post-CBT testing mode preference and mean of CBT scores.

Another step analysis of the results of the study was to examine whether the examiners have performed better performance of their preferred test versions depending on pre and post-CBT testing performance and its relationship to testing performance. The findings in (Table 8) showed that, those of CBT participants who preferred PPT version of the test (PPT performance, M=51.69) outperformed on CBT (M=66.11) and those who preferred CBT (PPT performance, M=50.18) performed better on PPT (M=59.41). While PPT participants who preferred PPT version of the test (PPT performance, M=50.32) in the PPT session outperformed on CBT (M=53.44) and those who preferred CBT version of the test (PPT performance, (M=51.63) performed better on PPT (CBT performance, M=47.76), and those who did not mind taking the test on either version, did better on CBT (M=54.46).

The findings showed that testing performance and testing mode preference of test takers had no positive interaction values, which means that testing mode preference inability to detect or influence the characteristics of the psychometric test, especially the validity of the test, the influence of exposure to the CBT version of the test on participants' posterior testing mode preference was examined.

Table 8. Descriptive statistics

PPTs	Paper	75	50.32	53.44	16.74	28.20
	No difference	12	48.18	54.46	11.77	15.96
	Onscreen	33	51.63	47.76	26.89	17.94
CBTs	Paper	18	51.69	66.11	14.33	38.43
	No difference	14	46.87	52.88	15.45	15.66
	Onscreen	88	59.41	50.18	19.35	19.35

The relationship of pre-CBT testing mode preference of different preference groups with their testing performances

**Note: Pr-CBT p refer to Pre-CBT performance and Po-CBT p refer to Post-CBT performance*

To show the difference between testing mode preference before and after exposure to CBT, the answers of the participants to the testing mode preference question were collected to show proportion responses, (Table 9) values indicted that On-paper (Pre-CBT) PPT (n=75, P=625) while (Post-CBT) CBT (n=18, P= 15) however, no difference (Pre-CBT) PPT (n=12, P=10), while (Post-CBT) CBT (n=14, P=11.66), but for the On-screen (Pre-CBT) PPT (n= 33, P= 275), while (Post-CBT) CBT (n= 88, P= 73). Findings revealed that although test takers show high preference for taking CBT, they did better on PPT version of the test. We find that the number of participants who preferred to take PPT by reviewing these values from the results and those participants who preferred to take the test in either version changed for the side of the participants who preferred to take CBT.

Table 9. Descriptive statistics

Preferred testing	(Pre-CBT) PPT		(Post-CBT) CBT	
Mode	Frequency	Percentage	Frequency	Percentage
On paper	75	62.5	18	15
No difference	12	10	14	11.66
Onscreen	33	27.5	88	73
Total	120	100	120	100

Differences between pre and post-CBT testing mode preferences

From (Table 9) we observe that: 62.5%, 27.5% of participants preferred to take PPT and CBT versions of the test, respectively, before the exposure to the CBT. Besides, 10% of participants didn't mind taking the test in either mode. After implementing CBT version of the test, only 15% still preferred to take PPT and 11.66% of the participants didn't mind taking the test in either mode. In this step of the study, the greatest percentage (73) was provided by the participants who chose CBT version of the test. The findings revealed that, after exposure to the CBT, the number of participants who preferred to take PPT and those participants who preferred to take the test in either mode changed in favor of the participants who preferred to take CBT.

4. CONCLUSION:

The present study was conducted for the purpose of investigating and determining whether there were any statistically significant differences in the scores of subjects obtained from the application of the compulsory islamic culture course test on the students of the University of Jordan and on the CBT and PPT versions. The results of the statistical analysis of the differences between females and males in performance on the test of the CBT version, indicated that there were no significant differences between the sexes in relation to there scores through the Two versions in the current study, it was found that sex differences were not a factor with a clear and strong performance on the subjects of both sexes effect.

This outcome is inconsistent with the findings of some studies of the no correlation indicator or a low correlation indicator either on the pre-CBT or post-CBT studies Such as (Flowers et al., 2011), (Higgins et al., 2005) and (Khoshsima et al., 2017). It is clear from the results of the present study, although the test takers CBT version may change its preference for the pre-test version, which may lead to acceptable performance relative to the type of test version, preferring the type of pre-test version as a moderate variable does not have that strong or influential effect on the examiners performance of the CBT version. The present study recommends further research and studies on the same subject taking into account the specialty of the examine, test anxiety, the number of test items, the test time implementing, and the cultural background of the examine, further replications of the study with more participants who are less homogeneous would be desirable thereafter. Conduct further studies to see if the tests give similar grades when administered in PPT or CBT forms. Furthermore, by examining item-level performance in addition to the performance of the test level, this study provided an opportunity to review differences in form at the item level.

Declaration:

Ethics approval and consent to participate: The study was conducted in accordance with the Declaration of Helsinki, and all participants provided written, informed consent to participate.

Consent for publication: There are no details on individuals reported within the manuscript.

Competing interests: The author declares that he has no competing interests.

Funding: The research was funded personally by the researcher

Availability of data and materials: Data and materials described in the manuscript are available by contacting the author of the article at (abd_275@yahoo.com).

Acknowledgments

The cooperation of the University of Jordan, represented by its various faculties, especially the faculties of pharmacy, science and Sharia, facilitated the study by providing all data and information and facilitating the task of conducting the tests related to the study, as well as thanks to the Rapid Technical Center in Amman –Jordan represented by Mr. Abdul Mahdi Ali Al-Akayleh who provided all the technical and advisory capabilities Current study tools.

5. REFERENCES:

Ackerman, R, & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology*, 17(1), 18-32.

Al-Amri, S. (2009). *Computer-based testing vs. paper-based testing: establishing the comparability of reading tests through the evolution of a new comparability version in a Saudi EFL context*. (Unpublished doctoral dissertation), University of Essex, England.

Anne, A., Walgermo, B., & Bronnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, 58 (2), 61-68.

Aslan, O. (2006). New way of learning: E-Learning, Fırat University, *Journal of Social Science*, (16) 2, 121-131.

Chapelle, C. A, & Douglas, D. (2006). *Assessing language through computer technology*. New York: Cambridge University Press.

Clariana, R, & Wallace, P. (2005). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, 33 (5), 593-602.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial/ethnic and gender groups. *Journal of Educational Measurement*, 39 (1), 133-147.

Creed, A., Dennis, I., & Newstead, S. (1987). Proof-reading on VDUs. *Behaviour & Information Technology*, 6 (1), 3–13.

Destefano, D, & Lefevre, J. (2007). Cognitive load in hypertext reading: *A review*. *Computers in Human Behaviour*. (23) 3, 1616–1641.

Dillon, A. (1994). Designing usable electronic text: *Ergonomic aspects of human information usage*. London: Taylor & Francis.

Dundar, H, & Akcayır, M. (2012). Tablet vs. Paper: The effect on learners' reading erformance. *International Electronic Journal of Elementary Education*, 4 (3), 441-450.

Ekrem, S. (2014). Computer versus Paper-Based Reading: A Case Study in English Language Teaching Context. *Mevlana International Journal of Education (MIJE* 4 (1).

Evans, J. D. (1996). Straightforward statistics for the behavioral sciences. *Pacific Grove*, CA: Brooks/Cole Publishing.

Fuhrer, S. (1973). A comparison of a computer-assisted testing procedure and standardized testing as predictors of success in community college technical mathematics (Doctoral dissertation), *New York University, 1973)*. Dissertation Abstracts International, 34 (6), 3086.

Fleming, S., Hiple, D., & Du, Y. (2002). Foreign language distance education at the University of Hawaii. In C. A. Spreen (ed.), *New technologies and language learning: issues and options (Technical Report #25)* Honolulu, HI: University of Hawaii, Second Language Teaching and Curriculum Center, 13-54.

Flowers, C., Do-Hong, K., Lewis, P., & Davis, V. C. (2011). A comparison of computer-based testing and pencil-and-paper testing for students with a read- aloud accommodation. *Journal of Special Education Technology*, 26 (1), 1-12.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups M. *Journal of Educational Measurement*, 39 (2), 133-147.

H, Khoshsima, M., Hosseini, S., & Hashemi, A. (2017). Cross-Mode Comparability of

Computer-Based Testing (CBT) Versus Paper-Pencil Based Testing (PPT): An Investigation of Testing Administration Mode among Iranian Intermediate EFL LearnersToroujeni1. *English Language Teaching*, 10 (2), 64-72.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment*, 3 (4), 1-36.

Holtzman, W. H. (1970). *Individually tailored testing: Discussion. In W. H. Holtzman (Ed.),* Computer-assisted instruction, testing, and guidance. New York: Harper & Row.

Kate Tzu, C. C. (2012). Elementary EFL teachers' computer phobia and computer self-efficacy in Taiwan. TOJET: *The Turkish Online Journal of Educational Technology*, 11 (2), 100-107.

Kim, J. (2013). Reading from an LCD monitor versus paper: Teenagers' reading performance. *International Journal of Research Studies in Educational Technology*, (2) 1, 15-24.

Lottridge, S.M., Nicewander, W.A., Schulz, E.M., & Mitzel, H.C. (2008). *Comparability of paper-based and computer-based Tests*: A review of the methodology. Monterey, CA: Pacific Metrics Corporation, Section 2: Studies of Comparability Methods, 13-32.

Makiney, J. D., Rosen, C., Davis, B. W., Tinios, K., & Young, P. (2003). Examining the measurement equivalence of paper and computerized job analyses scales. *18th Annual Conference of the Society for Industrial and Organizational Psychology*, Orlando, FL.

Mojarrad, H., Hemmati, F., Jafari Gohar, M.,& Sadeghi, A. (2013). Computer-based assessment (CBA) vs. Paper/pencil-based assessment (PPBA): An investigation into the performance and attitude of Iranian EFL learners' reading comprehension, *International Journal of Language Learning and Applied Linguistics World*, 4 (4), 418-428.

Niemeyer, C. (1999). A Computerized Final Exam for a Library Skills Course. *Reference Services Review*, 27 (1), 90-106.

Pinsoneault, T. B. (1996). Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota Multiphasic Personality Inventory-2. *Computers in Human Behavior*, 12, 291-300.

Poggio, J., Glasnapp, D., Yang, X.,& Poggio, A. (2005). A Comparative Evaluation of Score Results from Computerised and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program. *The Journal of Technology Learning and Assessment*, 3 (6), 1-31.

Pommerich, M. (2004).Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *The Journal of Technology Learning, and Assessment*, 2 (6), 1-45.

Salimi, H., Rashidy, A., Salimi, H.,& Amini, M. (2011). Digitized and non-Digitized Language Assessment: A Comparative Study of Iranian EFL Language Learners. International Conference on Languages, Literature and Linguistics IPEDR vol.26. IACSIT Press, Singapore.

Woolfolk, A. (2007). *Educational psychology* (10th ed.).New York: Pearson Education, Inc.

Zandvliet, D., & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education*, 29 (4), 423-438.

International Journal of
Assessment Tools in Education

Volume: 5 Number: 1
January 2018

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Development of the rubric self-efficacy scale

Perihan Güneş, Özen Yıldırım, Miraç Yılmaz

To cite this article: Güneş, P., Yıldırım, Ö., & Yılmaz, M. (2018). Development of the rubric self-efficacy scale. *International Journal of Assessment Tools in Education*, 5(1), 187-200. **DOI: 10.21449/ijate.373040**

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 187-200

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Development of the rubric self-efficacy scale

Perihan Güneş[1], Özen Yıldırım[2*], Miraç Yılmaz[3]

[1]Aksaray University, Education Faculty, Nigde, Turkey
[2]Pamukkale University, Education Faculty, Denizli, Turkey
[3]Hacettepe University, Education Faculty, Ankara, Turkey

Abstract: The purpose of this study is to develop a valid and reliable measurement tool determining teachers' self-efficacy regarding rubrics. Especially in educational environments, rubrics are measurement tools used in the assessment phase of student products usually based on higher-order thinking skills. Determination of teachers' self-efficacy regarding rubrics can give researchers an idea on how often and how accurately teachers use such tools. For this reason, the existence of a tool accurately measuring self-efficacy variable is necessary. This study's sample consists of 641 elementary, middle and high school teachers. To determine teachers' self-efficacy levels regarding rubrics, 47-item draft was developed. As a result of validity and reliability analyzes, a 28-item measurement tool with a four-factor structure was obtained. The total scale's and sub-factors' internal consistency is quite high. Using this scale, researchers can examine the relationships between teachers' self-efficacy and various variables that play an important role in education. In addition, comparative studies on the intended use of rubrics can be conducted by determining teachers' self-efficacy levels regarding rubrics.

ARTICLE HISTORY

Received: November 01, 2017
Revised: December 12, 2017
Accepted: December 28, 2017

KEYWORDS

Rubric, Teacher Efficacy Scale Development, Psychometric Properties, Performance Tasks

1. INTRODUCTION

The changes in social needs also bring about changes in the qualities people are required to have. In recent years, societies are in need of individuals who can analyze information, think creatively, impart the information they have learned into their daily lives and do research, and who have a developed critical perspective. Many countries have been constantly changing their curriculum to meet this need. The changes made are not only limited to the teaching approaches but also reflect on measurement and evaluation approaches. The question of how to evaluate these higher-order skills needed and the insufficiency of the available tools (oral exams, written exams, tests, etc.) led to complementary measurement and evaluation approaches, which enable these skills to concretize and thus to be measured, to take center stage. Complementary measurement and evaluation approaches provide performance-based assessments of the process in which the product was produced as well the product itself. Rubrics are one of the most common measurement tools used for this purpose.

*Corresponding Author E-mail: ozen19@gmail.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.373040**

Researchers define rubrics in different ways. However, according to the most commonly used definition, rubrics are tools that clearly specify the criteria which will be used to evaluate the observed performance, define the behaviors which the individuals have to exhibit in each criterion, and rank these performances from best to worst (or vice versa) (Andrea & Du, 2005; Andrade et al., 2009; Brookhart, 2013; Popham, 1997; Reddy & Andrade, 2010). Rubrics have three basic characteristics: evaluation criteria, criterion definitions, and scoring strategy (Popham, 2007). Evaluation criteria indicate according to which criteria a performance will be evaluated (Wiggins, 1991). Criterion definitions are detailed descriptions reflecting the performance levels of performance criteria scored from best to worst. Scoring strategies provide information on whether the scoring will be on the performance process or the product (Moskal, 2000).

In recent years, attempts to develop characteristics of higher-order thinking skills in schools and easier evaluation of products and process of these characteristics' popularized rubrics. Rubrics contribute significantly to both the teaching and evaluation process by presenting clear and well-defined criteria for the performance that needs to be exhibited. The most important characteristic of rubrics is that they clearly present teachers' learning objectives to the students. In addition, with the clear criteria presented in rubrics, teachers can provide students with detailed feedbacks about the products' weaknesses and strengths (Andrade, 2005). At the same time, detailed feedback mechanism supports the development of students' peer and self-evaluation skills (Panadero et al., 2016). Clear and well-defined criteria in rubrics allow the performance evaluation process to be transparent and consistent (Jonsson, 2014). This has a positive effect on the reliability of performance evaluation. Rubrics with well-defined performance criteria reduce the risk of different interpretation of the exhibited performance by evaluators (Reynolds et al., 2009) and the risk of incorrect scoring due to different interpretations (Venning & Buisman-Pijlman, 2013). In addition to these, rubrics support the development of psychological structures like self-efficacy and self-regulation which positively affect learning (Panadero & Jonsson, 2013).

Today, thanks to performance-based evaluations, teachers can easily evaluate whether students gained higher-order thinking skills or not at the end of their completed complex performance tasks (making presentation, designing model, writing an original story, etc.). For this reason, it is assumed that teachers have sufficient knowledge to use rubrics in educational settings and interpret the results, and they are expected to use these tools appropriately in schools. However, the studies conducted put forth that teachers have difficulties in how to prepare, implement and evaluate performance-based approaches and that they want to be informed on these issues (Metin 2013; Metin & Özmen, 2010). In this context, it is important to determine teachers' self-efficacy beliefs regarding rubrics, which are among the complementary measurement tools. Therefore, within the scope of this study, it was aimed to develop a tool measuring teachers' self-efficacy regarding rubrics.

Bandura(1977, 1994) defined the term 'self-efficacy', that he expressed as one of the most important factors that have an impact on the human behavior, as the self-belief of an individual in her/his competence or ability of successfully accomplishing a task. Bandura (1994) indicates that the beliefs on our abilities are influential on self-efficacy. The possession of a strong or a weak self-efficacy has an impact on the behavior or performance of an individual (Zimmerman, 2000). A strong self-efficacy belief is a behavior that increases the motivation of an individual with regards to overcoming a problem when a problem is confronted and enables an individual to put an effort. On the other hand, a weak self-efficacy belief prevents an individual to perform a task or finalize it (Jerusalem, 2002). A strong self-efficacy emotion is effected from the experience an individual had, other individuals' experiences, the expressions of an individual to perform a task, and from the emotional state

of an individual in the time that the behavior is displayed (Bandura, 1994). Schwarzer (1993) state that self-efficacy might be associated with various particular fields such as education, social, development and health. Moreover, Bandura (1977) remarked that individuals have different levels of self-efficacy in different fields, in other words, self-efficacy might alter according to the field and situation. For instance, an individual may have a high self-efficacy in a particular field, and low-efficacy in another field.

The belief of self-efficacy has been frequently used in the research studies related to learning and teaching (Özkan, Tekkaya & Çakıroğlu, 2002; Riggs & Enochs, 1990; Tschannen–Moran & Woolfolk–Hoy, 2001; Elias and Loomis, 2002). The self-efficacy of teachers, which is one of the most important factors in terms of learning and teaching also plays an important role. Teacher self-efficacy is the belief that teachers have about their abilities towards difficult or low-motivated students to participate in class and learn (Bandura, 1977). In the literature, there are several studies on teacher self-efficacy (Caprara, Barbaranelli, Steca, & Malone, 2006; Tschannen-Moran, Woolfolk-Hoy & Hoy, 1998; Tschannen–Moran & Woolfolk–Hoy, 2001; Yılmaz et.al 2004;). In this context, examination of the existed beliefs of teachers on applying subsidiary assessment and evaluation instruments. This situation might inform about how often and how correctly teachers use these instruments in in-class applications.

2. METHOD

This study is a scale development study using the basic survey model.

2.1. Study Group

This study was carried out during the 2016-2017 academic year. The scale development phase of the study was conducted with 641 elementary, middle and high school teacher who were knowledgeable about rubrics.

During the first phase of the study, the data obtained from 216 teachers were used in principal factor analysis and the data obtained from the remaining 425 teachers were used in confirmatory factor analysis. 327 (51%) of the participants were female and 314 (49%) were male. When the school levels were taken into consideration, the number of participant elementary school teachers (73.5%) were higher than the number of participant middle and high school teachers (26.5%). In order to increase the study impact, data from 16 different cities from Turkey's seven regions were collected. Convenience sampling method was used to reach the sample.

2.2. Data Collection

The validity and reliability works of the Rubric Self-Efficacy Scale was obtained at the end of the pilot study conducted on the selected sample.

2.2.1 Rubric self-efficacy scale

The self-efficacy scale regarding rubrics was developed similar to the scaling approach based on grading totals developed by Likert (1932). During the scale development, first, literature on self-efficacy was reviewed. As a result of the review, literatures on rubrics and self-efficacy were reached. When the literature was examined, it was seen that there was not a measurement tool determining *"teachers' rubric self-efficacy"* in Turkish or in another language. Therefore, no direct resource was used while developing the items. In addition the literature review, ten elementary and high school teachers were asked to explain their views on the preparation, implementation and evaluation of rubrics in the classroom and their positive or negative experiences with rubrics. Based on the qualitative data obtained, 47 items on teachers' preparation, implementation and evaluation of rubrics were developed.

During the scale's pilot study development phase, the items were examined by two measurement and evaluation experts and two Turkish language experts. According to the views taken from them, researchers removed 12 items from pilot study of scale form due to the fact that they did not reflect what they intended for and that they had ambiguities. The other items were organized according to the expert opinions. In pilot application, there were 20 positive statements putting forth teachers' high self-efficacy level regarding rubrics and 15 negative statements emphasizing teachers' low self-efficacy level. Teachers express how much they agree or disagree with the statements by choosing responses of Strongly Agree (5), Agree (4), Neither Agree or Disagree (3), Disagree (2) and Strongly Disagree (1).

2.3 Data Analysis

In the scale development phase, first, principal component analysis technique was used to put forth the state of the data structure and to reduce factor, and later confirmatory factor analysis was used to test the structure. Additionally to prove validity, item-total test correlation and a correlation coefficient from the upper and lower 27% of the total group was tested. Also for reliability testing Cronbach Alfa level of each factors was found.

Before principal component factor analysis, the suitability of the data structure for analysis was examined. Multivariate and univariate extreme values were identified, and 12 people were left out of the analysis because of the unexpected data structure. KMO (Kaiser-Meyer-Olkin) value determines how suited the data structure is for factor analysis based on the sampling adequacy, and Bartlett's Test of Sphericity informs about the state of multivariate normal distribution of the data. Table 1 presents the statistics regarding the KMO and Bartlett's Test of Sphericity.

Table 1. KMO and Bartlett's Test of Sphericity

KMO		.79
Barlet Test	Ki-square	2351.76
	df	59
	p	.00

When Table 1 is examined, the KMO value was found to be 0.79. According to this value, the sample size is at an adequate level to continue factor analysis. Whether the data set met the multivariate assumption or not was checked with Bartlett's Test of Sphericity. The value obtained show that data set met the assumption of multivariate normality ($\chi 2$= 2351.76; $p<0.01$). In addition to these assumptions, multicollinearity problem between the variables was examined with Pearson Product-Moment Correlation, and it was found that there was no multicolllinearity. During the principal component analysis, factors with Eingen values greater than 1 were taken into consideration, and items with a factor load of at least 0.32 (Tabachnick & Fidel, 2001) were accepted and selected for the real scale. Cronbach's Alpha value, which determines the internal consistency, that is, how closely correlated the items are with each other and the test, was examined for the internal consistency according to the total scale and sub-dimensions. For item discrimination index, the groups of the upper and lower 27% were compared, and item total test correlations were examined for validity testing.

After verify the scale's structure, confirmatory factor analysis method was used. This phase includes the testing process of the measurement model. By this means, whether the factorized structure is verified as a model or not with the principal component analysis was examined. Before starting the confirmatory factor analysis, the data structure of 425 people different than the principal component analysis was examined, and extreme and missing values were checked. Eight people were excluded from the analysis because of their unexpected data structure in terms of univariate and multivariate extreme values. When the missing data was

examined, it was determined that the missing data structure was 1.25%, and the researchers decided to assign missing data based on the mean. Since the data did not meet the assumption of normal distribution during the confirmatory factor analysis, the data were normalized, and the analysis continued. The confirmatory component analysis allowed that each observed variable showed relationship with only the latent variable under it.

3. FINDINGS

This section of the study includes findings regarding the principal component analysis, item-total test correlation, upper and lower %27 total group analysis, reliability analysis and confirmatory factor analysis of rubric development.

3.1. Principal component analysis of Rubric Self-Efficacy Scale

According to the principal component analysis method done to determine the scale's factor structures, nine factors with Eigen values higher than 1.00 were obtained. These nine factors reflect 65.17% of the total variance. Findings based on Eigen values and the variances they explain are given in Table 2.

Table 2. Eigen values and the explained variances

Factor	Eigen value	Variance %	Total Variance %
1	6.78	19.37	19.37
2	6.03	17.23	36.60
3	2.22	6.35	42.95
4	1.95	5.56	48.51
5	1.35	3.86	52.38
6	1.22	3.49	55.87
7	1.17	3.34	59.22
8	1.06	3.04	62.26
9	1.02	2.91	65.168

The first four factors explain the 48.51% of the total variance. After these four factors, the contribution of other factors on the percentage of the total variance decreases. It is seen that the four-factor structure adequately explain the studied variable. This is presented in the scree plot (Figure 1) showing eigenvalue components.

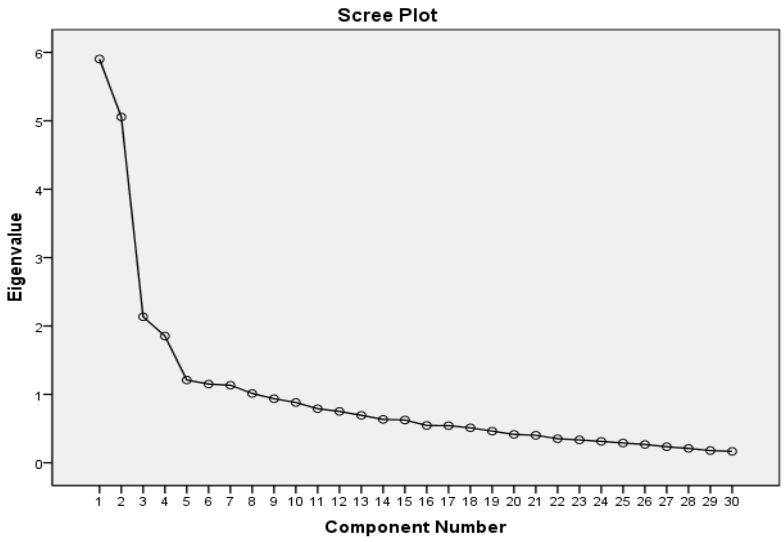


Figure 1. Scree plot

At the scree plot the slope with high acceleration where rapid decreases occur points out to a considerable amount of factor numbers. When Figure 1 is examined, it is seen that after four factors there is a routinized variation. After evaluating Table 2, Figure 1 and dimensions taken into consideration during item writing together, it was decided that the number of factors should be four. However, when the factor load values were examined before rotation, it was determined that factor load values of all items were greater than 0.32, and the smallest value was 0.486 and the greatest value was 0.76.

The step taken into account in the factorization process is the determination of the rotation method used. Varimax has been preferred as a rotation method since it was not expected that there would be a high degree of correlation among the factors that emerge in the principal component analysis. According to findings obtained from the rotation, 11 items under factor 1, nine items under factor 2, ten items under factor 3 and five items under factor 4 were determined. When the distribution of the items according to the factor load values is examined, the lowest load value is 0.43 and the highest load value is 0.79. When the items' cross loading is examined, it is seen that five items are collected under more than one factor and have a high loading value in each factor. The difference between factor load values is less than 0.10. Starting from the first item with the closest load value, the items were removed from the scale.

Table 3. Items' factor load values obtained as a result of factor analysis

Item	Factor 1 Loadings	Factor 2 Loadings	Factor 3 Loadings	Factor 4 Loadings
M19	0.75	0.19		-0.11
M8	0.72	0.20		
M9	0.67		0.27	
M20	0.65	0.22		
M13	0.62	0.19		0.11
M7	0.60	-0.12		
M10	0.59	0.22		
M27	0.59	0.28	-0.13	
M2	0.42	0.21	0.14	0.21
M39		0.78		
M41	0.19	0.77		
M33	0.14	0.70	-0.11	-0.22
M40	0.25	0.68	0.16	
M44	0.11	0.65		
M43	0.39	0.60	0.13	
M29	0.34	0.56	-0.11	
M32	0.18	0.41		-0.28
M12		0.16	0.77	
M15		-0.15	0.70	
M11	0.22		0.69	0.11
M6			0.66	0.18
M26			0.64	0.25
M18		-0.21	0.57	0.22
M24	0.14		0.55	0.26
M37			0.47	0.32
M31			0.14	0.82
M30	0.10		0.12	0.81
M28			0.20	0.79
M42			0.17	0.68
M35			0.22	0.66

The analyses were repeated in this order. Yet, since the cross loadings remained, these five items were not included in the scale. 30 items collected under four factors explain 49.05% of the total variance. The factor load values of the items that were collected under four factors as a result of factor analysis and were decided to be kept in the scale are shown in Table 3 below.

3.2 Item discrimination and examination of the test's reliability

In order to determine items' discrimination levels, item-total test correlation coefficients and item discrimination values for groups of the upper and lower 27% were examined. The findings are given in Table 4.

Table 4. Item analysis results

Item	Item-total test correlation N=204	Upper and lower 27% Nupper=Nlower=56
M2	0.478	-7.21***
M6	0.31	-3.68***
M7	0.32	-4.89***
M8	0.49	-7.99***
M9	0.45	-8.20***
M10	0.47	-7.19***
M11	0.32	-5.03***
M12	0.39	-5.92***
M13	0.32	-4.92***
M15	**0.10**	**-1.55**
M18	**0.19**	**-2.42***
M19	0.43	-6.19***
M20	0.34	-5.43***
M24	0.32	-5.60***
M26	0.34	-5.20***
M27	0.41	-6.59***
M28	0.56	-6.27***
M29	0.57	-7.18***
M30	0.53	-5.72***
M31	0.55	-6.36***
M32	0.38	-3.78***
M33	0.52	-5.79***
M35	0.33	-3.09***
M37	0.40	-4.57***
M39	0.57	-7.31***
M40	0.64	-8.91***
M41	0.61	-8.13***
M42	0.47	-4.20***
M43	0.63	-8.46***
M44	0.47	-5.36***

***p<0.001 *p<0.05

When item-total test correlations explaining the relationship between the scores from the items and the total score of the scale are examined in determining the item discrimination levels, it is seen that the correlation of item 15 and item 18 with the total has the lowest correlation scores, 0.102 and 0.197 respectively. Item-total test correlation values of the other items range between 0.65 and 0.31. On the other hand, when the difference between item scores' means among groups of the upper and lower 27% were examined, it was found that item 15 was not discriminative and for item 18 the mean difference between the lower and

upper groups is at a significant level of 0.05, but it was close to each other. Based on the two different discrimination findings, item 15 and 18 were excluded from the scale. When the reliability of the rest 28-item test is examined, the Cronbach's alpha value was determined as 0.85. This indicates that the test measures with high reliability. Reliability factor has also been tested for each factor. While the Cronbach's alpha value of the first factor was 0.80, the Cronbach's alpha value of the second factor was 0.89. The Cronbach's alpha value of the third factor was 0.70 whereas the Cronbach's alpha value of the fourth factor was 0.83. These results showed that sub-dimensions had high reliability (internal consistency).

As a result of the principal components analysis, 28-item scale is grouped under four factors. There are nine items under first factor, eight under the second, six under the third and five under the fourth factor. Factors were named as followed: Factor1 Efficacy of monitoring student development, Factor2 Efficacy of monitoring teaching, Factor3 Efficacy to overcome learning environment difficulties, Factor3 Efficacy of rubric preparation

3.3. Confirmatory factor analysis of the rubric self-efficacy scale

In the third phase of the study, structural validity of the items that were reduced to four factors through principal component analysis was tested with the confirmatory factor analysis. Table 5 was developed based on the results of the measurement model.

In Table 5, standardized loadings provide information about the correlation between the each observed variable and the latent variable that it is related to. While M8 (0.72) shows the highest correlation with Factor 1, M7 (0.47) shows the lowest correlation. Variability in Factor 1 is explained the most by the M8 (R^2=0.52) variable. M41 shows the highest correlation with Factor 2 whereas M32 shows the lowest correlation with Factor 2. For this reason, the variable with the highest R^2 coefficient is the M41(0.48) variable. When Factor 3 is examined, it is determined that M11 (0.65) shows the highest correlation with Factor 3 whereas M37 (0.38) shows the lowest. Most of the variability ($R^{2=}$0.42) in Factor 3 is explained by M11.When Factor 4 is examined, it is seen that M30 has the highest (0.79) correlation coefficient and M11 has the lowest (0.52). In this factor, most of the variance is explained by the M30 (R^2= 0.62) variable. When the error variances of the observed variables in the measurement model are examined, it found that error variances changed between 0.37 and 0,87. However, observed t values are calculated as significant for all variables (p<0.00). According to the findings obtained from the first examination of the measurement model, nothing disturbs the model's fit.

As a second examination, goodness of fit indices obtained from the measurement model were checked. Table 6 provides information on goodness of fit indices. When the significance of value X^2 that reveals the difference between the observed and expected matrices is examined, it was found that this value was significant (p<0.00). This can be due to the size sample size in the study. For this reason, the examination of goodness of fit indices continued. In terms of model goodness of fit indices, the indices other than GFI and AGFI have a good fit value whereas GFI (0.88) value and AGFI (0.86) value show a weak fit.

Table 5. Results of the measurement model

Factor/Item	Standardized Loadings	t-value	R^2
Factor1			
M2	0.52	10.61	0.27
M7	**0.47**	**9.54**	**0.22**
M8	**0.72**	**16.04**	**0.52**
M9	0.65	14.04	0.42
M10	0.57	11.97	0.32
M13	0.53	10.93	0.28
M19	0.70	15.47	0.49
M20	0.69	15.24	0.48
M27	0.58	12.25	0.34
Factor2			
M29	0.58	12.15	0.34
M32	**0.36**	**7.07**	**0.13**
M33	0.56	11.45	0.31
M39	0.70	15.38	0.49
M40	0.67	14.34	0.45
M41	**0.69**	**14.89**	**0.48**
M43	0.63	13.46	0.40
M44	0.61	13.93	0.37
Factor3			
M6	0.58	12.86	0.34
M11	**0.65**	**12.98**	**0.42**
M12	0.58	11.51	0.34
M24	0.47	8.92	0.22
M26	0.57	11.25	0.32
M37	**0.38**	**7.08**	**0.14**
Factor4			
M28	0.71	15.47	0.50
M30	**0.79**	**17.89**	**0.62**
M31	0.75	16.73	0.56
M35	0.55	11.34	0.30
M42	**0.52**	**10.49**	**0.27**

Table 6. Goodness of fit indices for measurement model

Fit indicates	Criteria	Value	Goodness
X^2 /sd	≤ 3	811.04 / 344 =	Good
RMSEA	0.05 ≤RMSEA≤0.08	0.057	Good
SRMR	0.05 ≤SRMR≤0.10	0.060	Good
NFI	≤0.90	0.91	Good
NNFI	≤0.90	0.94	Good
CFI	≤0.90	0.94	Good
GFI	≤0.90	0.88	Weak
AGFI	≤0.90	0.86	Weak
CN	≥ 200.00	216.29	Convenient sample

In addition to the goodness of fit, modifications suggested by the analysis outputs were checked. In modifications, it is suggested that M19 in Factor 1 is mapped to Factor 3 and factor 4, and M37 to Factor 2 and Factor 4. In the order of adjustments, M37 that lowered the X^2 value the most and then the exclusion of the item M19 from the model were examined. According to the findings, while there was a slight improvement only in the X^2/sd value, there was no improvement in the other goodness of fit indices AGFI and GFI value did not reach the desired fit level. For this reason, researchers decided that both of the items would remain in the model. No modifications were made during the confirmatory factor analysis.

4. DISCUSSION AND CONCLUSION

Many countries have been basing their educational programs on the constructivist approach, and they consider the use of process-based complementary measurement and evaluation tools like portfolios, projects, performance tasks and concept maps in addition to the use of traditional product-based measurement and evaluation tools like paper-and-pencil tests including open-ended, short and multiple choice questions as significant. In order for these complementary measurement and evaluation tools to be used effectively and efficiently, teachers must have knowledge and full competence in this area. For the determination of knowledge and self-efficacies, measurement tools that measure these qualities are also needed. By this means, more objective results can be reached. Within the scope of this study, a scale determining teachers' self-efficacy beliefs regarding rubrics was developed.

First, content validity of the scale developed by taking into the validity and reliability analysis was reached by taking the opinions of experts. Content validity is one of the leading validity analyses giving information on whether a scale measures based on an intended characteristic or not (Cronbach, 1990). Principal component analysis was done to determine the structure of the scale, and it was determined that the self-efficacy structure was grouped under four factors and these factors clearly explain nearly 50% of the variable. In multivariate designs, explained variance ratio is expected to be 40.00% and 60% (Scherer, Wiebe, Luther & Adam, 1998 cited in Tavşancıl, 2005). The obtained findings support the aforementioned resources. Moreover, when the slopes of the scree plot were examined, the four-factor structure can clearly be seen. Giving information about the internal consistency coefficient of the scale, Cronbach's alpha values show high reliability for both the scale itself and its sub-dimensions. This indicates that the items on the scale have a high correlation with each other. Cronbach (1970) stated that the scale would have high internal consistency if the alpha level on the scale is greater than 0.70.

According to the discrimination of the groups of the upper and lower 27% done to determine the item discrimination index levels, it was found that only two items did not differentiate between the positive and negative attitudes. This was determined by low correlation coefficient number, and they were not included in the scale. It is recommended that if the value of discrimination is lower than 0.19, the items should be revised, and if they cannot be adjusted, they should not be included in the scale (Kelley, 1939). The validity of the items was also examined by item-total test correlation. A low correlation suggests that the item should be removed from the scale (Cureton, 1966; Guilford, 1953). According to this finding, the same two items had the low correlation and they were removed from the scale.

Finally, the scale's structure validity was examined with confirmatory factor analysis, and it was determined that the structure put forth in the principal components analysis was reached again. When the analysis results were examined, it was seen that the chi square value was significant. Since the chi square value is affected by the sample (Byrne, 2003), goodness of fit indices was checked. The indices other than GFI and AGFI have a good fit value whereas GFI value and AGFI value show a weak fit. When the literature is examined, an AGFI value

greater than 0,85 can be considered as a good fit (Raykov & Marcoulides, 2006; Schermelleh-Engel Moosbrugger & Müller, 2003; Vieira, 2011).

The developed scale can be used for the purposes of related researches and institutions. Particularly, researchers working on complementary measurement and evaluation methods can examine the relationships between teachers' self-efficacy regarding rubrics and student performance according to different variables.

Acknowledgement

This work was supported by Aksaray University Scientific Research Projects Coordination Unit. Project Number: 2015-096

5. REFERENCES

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching*. 53(1), 27-30

Andrade, H. G., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation*, 10(3). Retrieved from http://PAREonline.net/getvn.asp?v=10&n=3

Andrade, H., Wang, X., Du, Y., & Akawi, R. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *The Journal of Educational Research*, 102(4), 287-302.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, 84 (2), 191-215.

Bandura, A. (1994*). Self-efficacy*. In V.S. Ramachaudran (Ed.), Encyclopedia of Human Behavior, (4), 71-81. New York: Academic Press.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading.* Alexandria, VA: ASCD.

Byrne, B. M. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, 34(2), 155-175.

Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: a study at the school level. *Journal of School Psychology*, 44, 473-490.

Cronbach, L. J. (1970). *Essentials of psychological testing*. 3rd Edition , Harper & Row.

Cureton, E. E. (1966). Corrected item-test correlations. *Psychometrik*a, 31, 93-96.

Elias, S., & Loomis, R. (2002). Utilizing need for cognition and perceived self-efficacy to predict academic performance. *Journal of Applied Social Psychology*, 32(8), 1687- 1702.

Guilford, J. P. (1953). The correlation of an item with a composite of the remaining items in a test. *Educational and Psychological Measurement*, 13, 87-93

Jerusalem, M. (2002). Theroretischer Teil - Einleitung I, *Zeitschrift für Pädagogik*, 44, 8-12

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education*, 39 (7), 840-852. doi:10.1080/02602938.2013.875117

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30(1), 17-24.

Likert, R. (1932). A Techniques for the measurement of attitudes. *Archives of Psychology*, 140, 5-53

Metin, M. (2013). Öğretmenlerin performans görevlerini hazırlarken ve uygularken karşılaştığı sorunlar. *Kuram ve Uygulamada Eğitim Bilimleri*, 13(3), 1645- 1673.

Metin, M., & Özmen, H. (2010). Fen ve teknoloji öğretmenlerinin performans değerlendirmeye yönelik hizmet içi eğitim (HİE) ihtiyaçlarının belirlenmesi. *Kastamonu Eğitim Dergisi*, 18(3), 819-838.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation*, 7(3),1-5

Özkan, Ö., Tekkaya, C., & Çakıroğlu, J. (2002). Fen bilgisi aday öğretmenlerin fen kavramlarının anlama düzeyleri, fen öğretimine yönelik tutum ve öz-yeterlik inançları. V. Ulusal Fen Bilimleri Eğitimi Kongresi, ODTÜ, Ankara.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144. doi:10.1016/j.edurev.2013.01.002.

Panadero, E., Jonsson, A., & Strijbos, J. W. (2016). *Scaffolding self-regulated learning through selfassessment and peer assessment: Guidelines for classroom implementation*. In D. Laveault & L. Allal (Eds.), Assessment for learning: Meeting the challenge of implementation (pp. 311–326). Cham: Springer International Publishing.

Popham, W. J. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership*, 55(2), 72-75.

Popham, W. J. (2007). *Classroom Assessment: What Teachers Need to Know*. Pearson Education, 5th Edition, USA.

Raykov T & Marcoulides G. A. (2006). *Fundamentals of structural equation modeling. A first course in structural equation modeling*. 2nd ed. London: Lawrence Erlbaum Associates; p.1-3, 41-3

Reddy, Y., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35, 435- 448. doi:10.1080/02602930902862859.

Reynolds, J., Smith, R., Moskovitz, C., & Sayle, A. (2009). BioTAP: A Systematic Approach to Teaching Scientific Writing and Evaluating Undergraduate Theses. *BioScience*, 59 (10), 896–903. doi:10.1025/bio.2009.59.10.11

Riggs, I. M. ve Enochs L. G. (1990). Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education*, 74(6), 625-637.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74.

Schwarzer R. (1993). General perceived self-efficacy in 14 cultures. Retrieved on 5-June 2007, at URL http://Web.Fu-Berlin.De/Gesund/Publicat/Ehpscd/Health/World14.Htm

Tabachnick, B. G., & Fidel, L. S. (2001). *Using multivariate statistics*. Fourth Edition. Needham Heights, MA: Allyn & Bacon.

Tavşancıl, E. (2005). *Tutumların ölçülmesi ve SPSS ile veri analizi*. Nobel Yayınları, Ankara

Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, 68, 202-248.

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing and elusive construct. *Teaching and Teacher Education*, 17, 783-805.

Venning, J., & F. Buisman-Pijlman (2013). Integrating assessment matrices in feedback loops to promote research skill development in postgraduate research projects. *Assessment and Evaluation in Higher Education*, 38 (5): 567–579.

Vieira, A. L. (2011). *Preparation of the analysis. Interactive LISREL in practice*. 1st ed. London: Springer.

Wiggins, G. (1991). Standart, not standardization: Evoking quality student work. *Educational Leadership.* 48(5), 18-25.

Yılmaz, M., Köseoğlu, P., Gerçek, C., Soran, H. (2004). Yabancı dilde hazırlanan bir öğretmen öz-yeterlik ölçeğinin Türkçeye uyarlanması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 27, 260-267.

Zimmerman, B.J. (2000). Self-Efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25, 82–91. doi:10.1006/ceps.1999.1016

Table Appendix 1. **Rubric Self-Efficacy Scale**

Rubric Self-Efficacy Scale		**Strongly Disagree**	**Disagree**	**Neither Agree or Disagree**	**Agree**	**Strongly Agree**
1	I believe, I have sufficient information on how to use rubrics in the classroom.					
2	I may have difficulties while preparing rubrics despite my experience in teaching.					
3	I believe use rubrics even if students have negative attitudes towards them.					
4	I can easily evaluate student performances with rubrics.					
5	I believe I have sufficient information on how to prepare rubrics.					
6	I can do applications using rubrics even if the students have not used them before.					
7	I may have difficulties while preparing rubrics even if I have theoretical knowledge.					
8	I may have difficulties while doing rubric applications because it takes a lot of time.					
9	If I encounter a problem while preparing a rubric, I can overcome it.					
10	I can increase student achievement by preparing effective rubrics.					
11	I can use rubrics effectively in group work.					
12	I may have difficulty in explaining the purpose of using rubrics to the students.					
13	I may have difficulty in adapting a rubric I have found from other sources to the subject.					
14	I can increase student interest towards the subject by rubric applications.					
15	I may have difficulty in scoring according to different rubric types.					
16	With rubrics, I can easily determine students' shortcomings during the learning process.					
17	I may have difficulty in determining the rubric type appropriate to the subject matter.					
18	While preparing the rubric, I may have difficulty in determining the learning objectives.					
19	Even if I have challenges in classroom management, I can do rubric applications.					
20	I can fairly evaluate the written exams with a rubric.					
21	While preparing the rubric, I may have difficulty in deciding on the behaviors to be measured.					
22	I may have difficulty in doing rubric applications in crowded classrooms.					
23	I believe I can improve myself in preparing rubrics by using different sources.					
24	I believe the students can easily understand the rubrics I prepare.					
25	With rubric applications, I believe I can lessen students' anxieties about learning the subject.					
26	Even if I try, while preparing rubrics, I may have difficulty in coming up with detailed definitions measuring student behaviors.					
27	I can prepare rubrics that would make students come up with quality works.					
28	If I try, I can get students gain the ability to use rubrics.					

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 187-200

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Turkish version *Research Article*

Dereceli Puanlama Anahtarına Yönelik Öz-Yeterlilik Ölçeğinin Geliştirilmesi

Perihan Güneş[1] **Özen Yıldırım**[*2] **Miraç Yılmaz**[3]

[1]Aksaray Üniversitesi, Eğitim Fakültesi, Türkiye
[2]Pamukkale Üniversitesi, Eğitim Fakültesi, Denizli, Türkiye
[3]Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara, Türkiye

Öz: Bu araştırmanın amacı, öğretmenlerin Dereceli Puanlama Anahtarları (DPA)'na yönelik öz yeterliklerini belirleyen geçerli ve güvenilir ölçme aracı geliştirmektir. Dereceli puanlama anahtarları özellikle eğitim ortamlarında öğrencilerin üst düzey zihinsel becerilerine dayalı ortaya koydukları ürünlerin değerlendirilmesi aşamasında kullanılan ölçme araçlarıdır. Öğretmenlerin DPA'ya yönelik öz yeterliklerinin belirlenmesi, onların bu tür araçları ne sıklıkla ve doğru olarak kullandıkları konusunda araştırmacılara bir fikir verebilir. Bu nedenle öz yeterlik değişkenini doğru olarak ölçen bir aracın varlığı gereklidir. Temel araştırma modeline dayalı olan bu araştırmanın örneklemini, ilköğretimde ve ortaöğretimde çalışan 641 öğretmen oluşturmaktadır. Öğretmenlerin dereceli puanlama anahtarına yönelik öz yeterlilik düzeylerini ortaya koymak için 47 maddeden oluşan taslak form hazırlanmıştır. Ölçeğin geçerliğini belirlemek için kapsam geçerliği, temel bileşenler analizi, madde toplam korelasyonu, alt-üst %27'lik gruplar için madde analizi ve doğrulayıcı faktör analizi yapılmıştır. Ölçeğin güvenirliği ise iç tutarlılık olarak incelenmiş ve Cronbach alfa değeri ile test edilmiştir. Analizler sonucunda 28 maddeden oluşan ve dört faktörlü bir yapıya sahip ölçek elde edilmiştir. Ölçeğin tamamının ve alt faktörlerinin iç tutarlılık katsayısı oldukça yüksektir. Bu ölçeğe dayalı olarak araştırmacılar öğretmenlerin öz yeterliği ile eğitimde önemli rol oynayan farklı değişkenler arasındaki ilişkiler incelenebilirler. Ayrıca öğretmenlerin DPA'ya yönelik öz yeterlilik düzeyleri belirlenerek DPA'ların kullanım amaçlarına dayalı karşılaştırmalı çalışmalar yapılabilir.

MAKALE SÜRECİ

Gönderim: 01 Kasım 2017
Düzeltme: 12 Aralık 2017
Kabul: 28 Aralık 2017

ANAHTAR KELIMELER

Rubrik, Öğretmen Özyeterliği, Ölçek Geliştirme,Psikometrik Özellikler, Performans Görevleri

1. GİRİŞ

Toplumsal ihtiyaçlardaki değişmeler ihtiyaç duyulan insan niteliklerinin de değişmesini beraberinde getirmektedir. Son yıllarda toplumlar bilgiyi analiz edebilen, yaratıcı düşünebilen, öğrendiği bilgileri günlük hayata aktarabilen, eleştirel bakış açısı gelişmiş, araştırma yapabilen bireylere daha çok ihtiyaç duymaktadırlar. Birçok ülke bu ihtiyacı karşılamak için öğretim programlarını sürekli değiştirmektedir. Yapılan değişiklikler sadece öğretim yaklaşımları ile

*Sorumlu Yazar E-mail: ozen19@gmail.com

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.373040**

sınırlı kalmayıp, ölçme değerlendirme yaklaşımlarına da yansımaktadır. İhtiyaç duyulan bu üst düzey becerilerin nasıl değerlendirileceği ile mevcut araçların (sözlü sınavlar, yazılı sınavlar, testler vb.) yetersiz kalması, bu becerilerin somut biçimde ifade edilmesini ve dolayısıyla ölçülmesini sağlayacak tamamlayıcı ölçme ve değerlendirme yaklaşımlarının ön plana çıkmasını sağlamıştır. Tamamlayıcı ölçme değerlendirme yaklaşımları, ürünün yanı sıra ürünün ortaya çıkma süreci hakkındaki öğretmenlere bilgi sunmaktadır. Örneğin öğrencinin kaynaklara nasıl ulaştığı, hangi kaynaklardan yararlandığı, bu kaynakları nasıl kullandığı hakkında ayrıntılı bilgilere ulaşılabilir. Dereceli puanlama anahtarı (DPA) bu amaçla kullanılan en yaygın ölçme araçlarından biridir.

Araştırmacılar tarafından DPA'lar çok farklı şekillerde tanımlanmakla birlikte en yaygın kullanılan tanım, gözlenen performansın hangi ölçütlere göre değerlendireceğini açıkça belirten, bireyin her bir ölçütte göstereceği davranışları tanımlayan ve bu performansları belirli değere göre sıralayan araçlardır (Andrea ve Du, 2005; Andrade ve diğerleri, 2009; Brookhart, 2013; Popham, 1997; Reddy ve Andrade, 2010).

DPA'lar değerlendirme ölçütleri, ölçüt tanımları ve puanlama stratejisi olmak üzere üç temel öğeye sahiptir (Popham, 2007). Değerlendirme ölçütleri, sergilenmesi gereken performansın hangi ölçütlere göre değerlendirileceğini göstermektedir (Wiggins, 1991). Ölçüt tanımları, performansta ele alınan her bir ölçüte dayalı performans düzeyleri de dikkate alınarak yazılmış detaylı tanımlardır. Puanlama stratejileri, puanlamanın sürece mi yoksa sonuca mı dönük yapılacağını hakkında bilgi vermektedir (Moskal, 2000).

Son yıllarda özellikle eğitim ortamlarında üst düzey zihinsel özelliklerin geliştirilmek istenmesi ve bu özelliklerinin ürünlerinin ve sürecinin DPA'lar sayesinde daha kolay değerlendirilebilmesi, bu ölçme araçlarını popüler hale getirmiştir. DPA'lar öğrenciden beklene performansla ilgili açık ve iyi tanımlanmış ölçütler sunarak hem öğretme hem de değerlendirme sürecine önemli katkılar sağlamaktadır. DPA'ların en önemli özelliği öğretmenlerin öğrenme ile ilgili hedeflerini öğrencilere açık bir şekilde sunmasıdır. Bunun yanı sıra DPA'larda sunulan açık ölçütler ile öğretmenler, öğrencilere çalışmalarının zayıf ve güçlü noktaları hakkında detaylı geri bildirimler verebilmektedir (Andrade, 2005). Aynı zamanda detaylı geribildirim mekanizması öğrencilerin akran ve öz değerlendirme becerilerinin gelişmesini desteklemektedir (Panadero ve diğerleri, 2016). DPA'larda ölçütlerin açık ve iyi tanımlanmış olması performans değerlendirme sürecinin şeffaf ve tutarlı olmasına olanak sağlamaktadır (Jonsson, 2014). Bu durum performans değerlendirmenin güvenirliği üzerinde pozitif etkiye sahiptir. İyi tanımmış performans ölçütlerine sahip DPA'lar sergilenen performansın değerlendiriciler tarafından farklı yorumlanma (Reynolds ve diğerleri, 2009) ve yorumlanma farklılığından kaynaklanan hatalı puanlama riskini düşürmektedir (Venning ve Buisman-Pijlman, 2013). Tüm bunların yanı sıra DPA'lar öğrenmeyi olumlu etkileyen öz yeterlilik ve öz düzenleme gibi psikolojik yapıların gelişmesini de desteklemektedir (Panadero ve Jonsson, 2013).

Günümüzde performansa dayalı değerlendirmeler sayesinde öğretmenler, öğrencilerin karmaşık performans görevleri (sunum yapma, model tasarlama, özgün bir hikâye yazma vb.) yerine getirerek üst düzey zihinsel becerileri kazanıp kazanmadığını rahatlıkla ölçebilir. Bu nedenle eğitim ortamlarında öğretmenlerin DPA'ları kullanma ve sonuçları yorumlama konusunda yeterli bilgiye sahip oldukları varsayılmakta ve bu araçları okullarda uygun bir şekilde kullanmaları beklenmektedir. Oysaki yapılan çalışmalar öğretmenlerin performansa dayalı yaklaşımların nasıl hazırlanacağı, uygulanacağı ve değerlendirileceği konusunda zorlandıklarını ve bu konularda bilgilenmek istediklerini ortaya koymaktadır (Metin ve Özmen, 2010; Metin 2013). Bu tür bir ölçme aracını kullanırken, bilgi eksiklikleri olan veya zorlanan öğretmenlerin, sınıf içi uygulamalarında bu araçtan yararlanma sıklıklarının da az olması beklendik bir durumdur. Bununla birlikte DPA'lar sadece öğretmenin öğrenci

performansı hakkında bilgi edinmesini değil, öğrencinin de görevine dayalı süreç ve üründe neleri yapıp yapamadığı hakkında bilgi edinmesini sağlar. Öğrenciler eksikliklerini görerek ilerideki görevlerinde bu eksikliklerinin üstesinden gelmeye çalışır. Bu bağlamda ilk olarak öğretmenlerin tamamlayıcı ölçme araçları içerisinde yer alan DPA'ya yönelik öz yeterlik inançlarının ortaya çıkarılması önemlidir. Ancak öğretmenlerin DPA'ya yönelik öz yeterliklerini ölçen geçerli ve güvenilir bir araç literatürde bulunmamaktadır. Bu nedenle araştırma kapsamında öğretmenlerin DPA'ya yönelik öz yeterliklerini ölçen bir araç geliştirmek amaçlanmıştır.

Bandura (1977, 1994) insan davranışlarını etkileyen önemli faktörlerden biri olarak ifade ettiği öz yeterliği, bireyin bir işi başarılı olarak yapıp yapmayacağı konusunda kendisine duyduğu inanç şeklinde tanımlamıştır. Bandura (1994), yeteneklerimize ilişkin var olan inançların özyeterlik üzerinde etkili olduğunu belirtmiştir. Bireyin güçlü ya da zayıf öz yeterlik inancına sahip olması, bireyin davranışı ya da performansı üzerinde etkilidir (Zimmerman, 2000). Güçlü öz yeterlik inancı, bireyin herhangi bir problemle karşılaştığında o problemle başa çıkacağına dair motivasyonunu artıran ve çaba göstermesini sağlayan bir davranıştır. Zayıf özyeterlik inancı ise bireyin bir işi yapması ya da sonuçlandırması için çaba göstermesine engel olmaktadır (Jerusalem, 2002). Güçlü bir öz yeterlik duygusu bireyin kendi yaşadığı deneyimlerden, başkalarının yaşadığı deneyimlerden, bireyin bir işi yapacağına dair motivasyon ifadeleri ile bireyin davranışı sergileyeceği andaki duygusal durumundan etkilenmektedir (Bandura, 1994). Schwarzer (1993) öz yeterliğin eğitim, sosyal, gelişim ve sağlık gibi pek çok özel alanla ilişkili olabileceğini bildirmektedir. Bandura ise (1977), bireylerin farklı alanlarda farklı öz yeterlilik düzeylerine sahip olabileceğini, yani öz yeterliğin alana ve duruma göre değişebildiğini dile getirmiştir. Örneğin bir birey bir alanda yüksek öz yeterliğe, başka bir alanda ise düşük öz yeterliğe sahip olabilir.

Öz yeterlik inancı, öğrenme-öğretme ile ilgili araştırmalarda sıklıkla kullanılmaktadır (Özkan, Tekkaya ve Çakıroğlu, 2002; Riggs ve Enochs,1990; Tschannen–Moran ve Woolfolk–Hoy, 2001; Elias ve Loomis, 2002). Öğretme ve öğrenme üzerinde en etkili faktörlerden biri olan öğretmenlerin, çeşitli alanlardaki öz yeterlikleri de oldukça önemlidir. Öğretmen öz yeterliği, bir öğretmenin zor ya da motive olabilecek öğrencilerin bile derse katılımını ve öğrenmesini sağlamaya yönelik yeteneklerine olan inancıdır (Bandura, 1977). Literatürde öğretmen öz yeterliği ilgi birçok çalışma yapılmıştır (Caprara, Barbaranelli, Steca ve Malone, 2006; Tschannen-Moran, Woolfolk-Hoy ve Hoy, 1998; Tschannen–Moran ve Woolfolk–Hoy, 2001; Yılmaz, Köseoğlu, Gerçek ve Soran 2004). Bu bağlamda öğretmenlerde tamamlayıcı ölçme değerlendirme araçlarını kullanabileceğine dair var olan inançların incelenmesi önem taşımaktadır. Bu durum öğretmenlerin sınıf içi uygulamalarında ne sıklıkla ve ne kadar doğru olarak bu araçları kullandıkları yönünde ön bilgi verebilir. Öğretmenlerin olumsuz öz yeterlik inançlarının belirlenmesi durumunda, DPA'nın kullanılmasını özendirecek ek bilgilendirme seminerleri düzenlenebilir.

2. YÖNTEM

Bu araştırma, temel tarama modelinde planlanmış bir ölçek geliştirme çalışmasıdır.

2.1. Çalışma Grubu

Bu çalışma, 2016-2017 eğitim öğretim yılında gerçekleşmiştir. Çalışmaya dereceli puanlama anahtarı hakkında bilgisi olan 641 ilköğretim ve ortaöğretim öğretmeni ile gerçekleştirilmiştir. Öğretmenlerin DPA'lar hakkındaki bilgisi kendi algılarına dayalı olarak yoklanmıştır. Araştırmanın birinci adımı kapsamında 216 öğretmenden elde edilen veriler temel bileşenler analizinde geriye kalan 425 öğretmenden elde edilen veriler ise doğrulayıcı faktör analizinde kullanılmıştır. Katılımcıların 327 (%51)'si kadın, 314 (%49)'ü ise

erkeklerden oluşmaktadır. Okul düzeylerine göre ilkokul düzeyinde örnekleme katılan öğretmen sayısı (%73,5), ortaöğretim düzeyindeki öğretmenlere (%26,5) göre daha fazladır. Çalışmanın yaygın etkisini arttırmak için Türkiye'nin yedi bölgesinin 16 farklı illindeki devlet okullarında görev yapan çeşitli branşlarda öğretmenlerden veriler toplanmıştır. Çalışma grubuna ulaşmada uygun örnekleme yöntemi kullanılmıştır.

2.2 Verilerin Toplanması

Dereceli puanlama anahtarına ilişkin öz yeterlik ölçeğinin geçerlik ve güvenirlik çalışmaları belirlenen örneklem üzerindeki pilot uygulama sonucunda elde edilmiştir. *Dereceli puanlama anahtarına ilişkin öz yeterlik ölçeği:* DPA'ya yönelik öz yeterlik ölçeği, Likert (1932) tarafından geliştirilen dereceleme toplamlarına dayalı ölçekleme yaklaşımının adımlarına benzer olarak geliştirilmiştir.

Ölçek geliştirilirken öncelikle, öz yeterliğe dayalı ilgili alan yazın taraması yapılmıştır. Tarama sonucunda DPA ve öz yeterliğe dayalı ilgili kaynaklara ulaşılmıştır. Literatür incelendiğinde "*DPA'ya yönelik öğretmenlerin öz yeterliklerini*" ortaya koyan Türkçe veya yabancı dilde bir ölçme aracına ulaşılmadığından maddelerin oluşturulmasında destekleyici doğrudan bir kaynaktan yararlanılmamıştır. Literatür taramasının yanı sıra ölçek maddelerinin oluşturulması amacıyla 10 kişilik ilköğretim ve 2 lise düzeyinde öğretmene DPA'ları sınıf içerisinde uygulama, hazırlama ve kullanmalarına yönelik görüşleri ve bununla birlikte DPA'ya yönelik varsa olumlu veya olumsuz deneyimlerini açıklamaları istenmiştir. Elde edilen nitel verilerden yararlanarak öğretmenlerin DPA'ları hazırlaması, kullanımı ve uygulamasına ilişkin öz yeterliklerini ortaya koyabilecekleri 47 madde oluşturulmuştur.

Ölçeğin pilot uygulamaya hazırlanması sürecinde maddelerin incelenmesini iki ölçme ve değerlendirme ve iki Türk dili uzmanı gerçekleştirmiş, alınan görüşlere göre 12 madde ölçülmek istenilen durumu yansıtmaması ve anlatım bozuklukların olması nedeniyle araştırmacılar tarafından ön uygulama formundan çıkarılmıştır. Diğer maddeler uzmanların görüşlerine göre düzenlenmiştir. Ön deneme uygulamasına hazır hale getirilen ölçekte öğretmenin DPA'ya yönelik öz yeterliğini ortaya koyan 20 olumlu cümle, öz yeterliğindeki düşüklüğü vurgulayan 15 olumsuz cümle yer almaktadır. Öğretmenler ifadelerin her birini katılıp katılmama durumuna göre '5' Tamamen Katılıyorum, '4' Katılıyorum, '3' Kararsızım, '2' Katılmıyorum ve '1' Hiç Katılmıyorum" biçiminde derecelendirilmiş seçeneklerden seçmektedir.

2.3. Verilerin Analizi

Ölçek geliştirme aşamasında ilk olarak veri yapısının durumunu ortaya koymak ve faktör indirgemek amacıyla temel bileşenler analizi tekniğinden yararlanılmış, daha sonra yapıyı test etmek için doğrulayıcı faktör analizi uygulanmıştır.

Temel bileşenler analizinden önce veri yapısının analiz için uygunluğu incelenmiştir. Çok değişkenli ve tek değişkenli uç değerler belirlenmiş, buna göre 12 kişi beklenmedik veri yapısına sahip olması nedeniyle analiz dışında bırakılmıştır. Test edilen diğer varsayımlar örneklem büyüklüğüne dayalı olarak veri yapısının faktör analizine uygunluğunu ortaya koyan KMO (Kaiser-Meyer-Olkin) değeri ve verilen çok değişkenli normal dağılım gösterme durumunu hakkında ipucu veren Barlett Testi'dir. Tablo 1'de KMO ve Barlett Küresellik Testine ilişkin istatistikler verilmektedir.

Tablo 1. KMO ve Barlett Testi

KMO		.79
Barlet Testi	Ki-kare	2351.76
	sd	595
	p	.00

Tablo.1 incelendiğinde KMO değerinin 0.79 olduğu gözlenmiştir. Bu değere göre örneklem büyüklüğü faktör analizine devam etmek için iyi düzeydedir. Veri setinin çok değişkenli normallik varsayımını karşılayıp karşılamadığı ipucu ise Barlett Küresellik Testi ile kontrol edilmiştir. Elde edilen değer, veri setinin çok değişkenli normallik varsayımını karşıladığını göstermektedir ($\chi 2$= 2351.76; $p<0.01$). Bu varsayımların yanı sıra değişkenler arasındaki çoklu bağlantı problemi de Pearson Momentler Çarpımı korelasyonu ile incelenmiş ve çoklu bağlantının olmadığı gözlenmiştir.

Temel bileşenler analizi sırasında faktör öz değeri 1'den büyük olan faktörler dikkate alınmış ve faktör yükleri en az 0.32 (Tabachnick ve Fidel, 2001) olan maddeler kabul edilerek asıl ölçek için seçilmiştir. Maddelerin birbirleri ve testle olan korelasyonunu veren iç tutarlılık anlamındaki Cronbach Alfa değeri testin tamamı ve alt boyutlarına göre güvenirlik için incelenmiştir. Maddelerin ayırt ediciliği için alt-üst %27'lik gruplar karşılaştırılmış ve madde toplam test korelasyonlarına bakılmıştır.

Ölçeğin yapısını doğrulamak için doğrulayıcı faktör analizi (DFA) yöntemi uygulanmıştır. Bu aşama ölçme modelinin test edilmesi sürecini içermektedir. Bu sayede temel bileşenler analizi ile faktörleştirilmiş yapının bir model olarak doğrulanıp doğrulanmadığı incelenmiştir. DFA'ya geçilmeden önce temel bileşenler analizinden farklı 425 kişilik veri yapısı incelenmiş, uç değerler ve kayıp değerlere bakılmıştır. Tek yönlü ve çok yönlü uç değerler bakımından 8 kişi beklenmedik veri yapısına sahip olması nedeniyle analiz dışında tutulmuştur. Verilerde kayıp veri durumu incelendiğinde kayıp veri yapısının %1.25'i kadar olduğu belirlenmiş ve araştırmacılar tarafından ortalamaya dayalı kayıp verinin atanması kararlaştırılmıştır. DFA sırasında verilerin normal dağılım varsayımını karşılamamasından dolayı veriler normalleştirilerek analize devam edilmiştir.

DFA sırasında Temel Bileşenler Analizinin sonuçları dikkate alınarak, her bir gözlenen değişkenin yalnızca kendi altında yer alan bir gizil değişkenle ilişki göstermesine izin verilmiştir.

3. BULGULAR

Çalışmanın bu bölümünde dereceli puanlama anahtarı geliştirmeye ilişkin temel bileşenler analizi, güvenirlik analizleri ve faktör analizine ilişkin bulgulara ver verilmiştir.

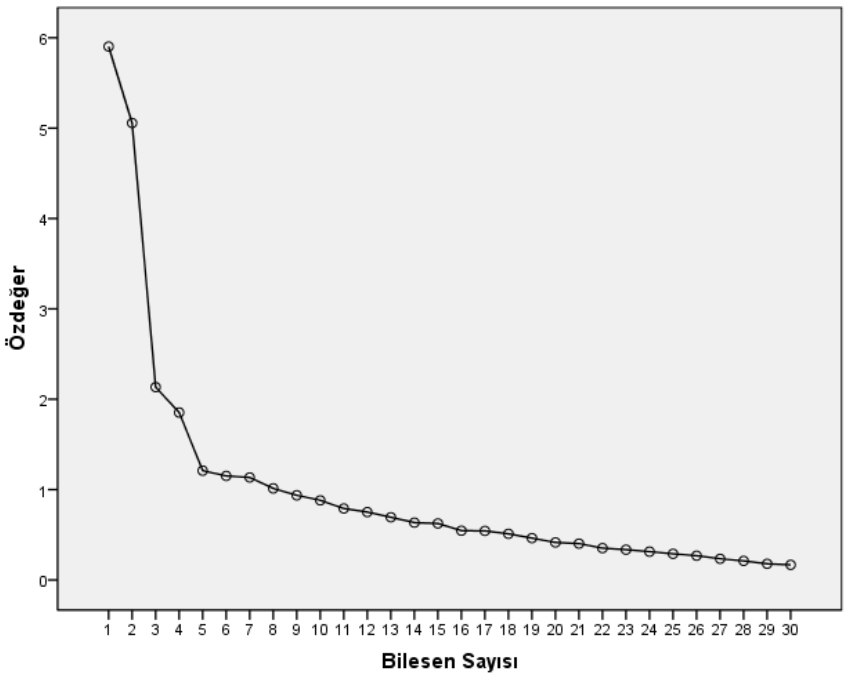
3.1. DPA öz yeterlik ölçeği temel bileşenler analizi

Ölçeğin faktör yapılarını belirlemek amacıyla yapılan temel bileşenler analiz yöntemine göre öz değeri 1.00'da yüksek dokuz faktör elde edilmiştir. Bu dokuz faktör toplam varyansın %65.168'ini yansıtmaktadır. Öz değer ve açıkladıkları varyanslara dayalı bulgular Tablo 2.'de verilmiştir.

Tablo 2. Öz değerler ve açıkladıkları varyanslar

Faktör	Öz değer	Varyans Yüzdesi	Toplam Varyans Yüzdesi
1	6.78	19.37	19.37
2	6.03	17.23	36.60
3	2.22	6.35	42.95
4	1.95	5.56	48.51
5	1.35	3.86	52.38
6	1.22	3.49	55.87
7	1.17	3.34	59.22
8	1.06	3.04	62.26
9	1.02	2.91	65.17

İlk dört faktör toplam varyansın %48.51'ini açıklamaktadır. Bu dört faktörden sonra diğer faktörlerin toplam varyans yüzdesine yaptığı katkı azalmaktadır. Dört faktörlü yapının araştırılan değişken için yeterli varyansı açıkladığı görülmektedir. Bu durum öz değer bileşenlerini gösteren çizgi grafiğinde de (Şekil 1) görülmektedir.



Şekil 1. Öz değer bileşen çizgi grafiği

Çizgi grafiğine göre yüksek ivmeli. hızlı düşüşlerin yaşandığı eğim önemli sayıda faktör sayısına işaret etmektedir. Şekil 1 incelendiğinde dört faktörden sonra eğim sabitlenmekte tek düze bir farklılaşma gözlenmektedir. Tablo 2, Şekil 1 ve madde yazımında dikkate alınan boyutlar bir arada değerlendirilerek faktör sayısının dört olmasına karar verilmiştir.

Bununla birlikte döndürmeden önce faktör yük değerleri incelendiğinde bütün maddelerin faktör yük değerlerinin 0.32'den büyük olduğu en küçük değerin 0.486 ve en büyük değerin 0.762 olduğu belirlenmiştir.

Faktörleşme sürecinde dikkate alınan adım kullanılan döndürme yönteminin belirlenmesidir. Temel bileşenler analizinde ortaya çıkan faktörler arasında yüksek derecede ilişki olması beklenmediğinden döndürme yöntemi olarak varimax tercih edilmiştir. Döndürme sonucunda elde edilen bulgulara göre 1. faktör altında 11 madde, 2. faktör altında 9 madde ve 3. faktör altında 10 madde ve 4. Faktör altında 5 madde belirlenmiştir. Maddelerin faktör yük değerlerine göre dağılımı incelendiğinde en düşük yük değeri 0.43 en yüksek yük değeri 0.79'tir. Bununla birlikte maddelerin binişiklik durumu incelendiğinde 5 maddenin birden fazla faktör altında toplandığı ve her faktörde de yüksek yük değerine sahip olduğu gözlenmiştir. Faktör yük değerleri arasındaki fark 0.10'dan küçüktür. En yakın yük değerine sahip ilk maddeden başlayarak sırasıyla maddeler ölçekten çıkarılmıştır. Analizler bu sırada tekrarlanmıştır. Ancak maddelerin binişikliklerinin kalkmamasından dolayı belirlenen bu beş madde ölçeğe dahil edilmemiştir.

Dört faktör altında toplanan 30 madde toplam varyansın %49.05'ini açıklamaktadır. Faktör analizi sonucunda dört faktör altında toplanan ve ölçekte kalmasına karar verilen maddelere ait faktör yük değerleri aşağıdaki Tablo 3'te gösterilmektedir.

Tablo 3. Faktör analizi sonucunda maddelere ilişkin elde edilen faktör yük değerleri

Madde No	Faktör 1 Yük Değeri	Faktör 2 Yük Değeri	Faktör 3 Yük Değeri	Faktör 4 Yük Değeri
M19	**0.75**	0.19		-0.11
M8	**0.72**	0.20		
M9	**0.67**		0.27	
M20	**0.65**	0.22		
M13	**0.62**	0.19		0.11
M7	**0.60**	-0.12		
M10	**0.59**	0.22		
M27	**0.59**	0.28	-0.13	
M2	**0.42**	0.21	0.14	0.21
M39		**0.78**		
M41	0.19	**0.77**		
M33	0.14	**0.70**	-0.11	-0.22
M40	0.25	**0.68**	0.16	
M44	0.11	**0.65**		
M43	0.39	**0.60**	0.13	
M29	0.34	**0.56**	-0.11	
M32	0.18	**0.41**		-0.28
M12		0.16	**0.77**	
M15		-0.15	**0.70**	
M11	0.22		**0.69**	0.11
M6			**0.66**	0.18
M26			**0.64**	0.25
M18		-0.21	**0.57**	0.22
M24	0.14		**0.55**	0.26
M37			**0.47**	0.32
M31			0.14	**0.82**
M30	0.10		0.12	**0.81**
M28			0.20	**0.79**
M42			0.17	**0.68**
M35			0.22	**0.66**

3.2. Madde ayırt ediciliği ve testin güvenirliğinin incelenmesi

Maddelerin ayırt edicilik düzeylerinin belirlenmesi amacıyla madde toplam test korelasyonu katsayıları ve %27'lik alt-üst gruplar için madde ayırt ediciliği değerleri incelenmiştir. Elde edilen bulgular Tablo 4'te verilmiştir.

Madde ayırt edicilik düzeylerinin belirlenmesinde maddelerden alınan puanlar ile ölçeğin toplam puanı arasındaki ilişkiyi açıklayan madde toplam test korelasyonları (madde ayırt edicilik değerleri) incelendiğinde; 15. ve 18. maddenin toplam ile korelasyonu sırasıyla 0.10 ve 0.19 olarak en düşük korelasyon değerine sahip olduğu gözlenmiştir. Diğer maddelerin madde toplam test koreslasyon değerleri 0.65 ve 0.31 arasında değişmektedir. Bununla birlikte alt üst yüzde 27'lik gruplar arası madde puanları ortalamaları arası farka bakıldığında 15. maddenin ayırt edici olmadığı. 18. madde de ise alt ve üst gruplar arası ortalama farkın 0.05 düzeyinde anlamlı olmasına rağmen değerin birbirine yakın olduğu gözlenmiştir. İki farklı ayırt edicilik bulgularına dayalı olarak 15. ve 18. maddeler ölçeğin dışında tutulmuştur.

Tablo 4. Madde analizi sonuçları

Madde No	Madde Toplam Test Korelasyonu n=204	T (Alt%27-Üst %27) nalt=nüst=56
M2	0.48	-7.20***
M6	0.31	-3.68***
M7	0.32	-4.88***
M8	0.50	-7.99***
M9	0.46	-8.20***
M10	0.47	-7.19***
M11	0.32	-5.03***
M12	0.39	-5.92***
M13	0.32	-4.91***
M15	**0.10**	**-1.55**
M18	**0.19**	**-2.41***
M19	0.42	-6.19***
M20	0.34	-5.433***
M24	0.32	-5.60***
M26	0.33	-5.20***
M27	0.41	-6.59***
M28	0.55	-6.26***
M29	0.57	-7.17***
M30	0.53	-5.72***
M31	0.54	-6.36***
M32	0.38	-3.78***
M33	0.51	-5.78***
M35	0.33	-3.08***
M37	0.40	-4.57***
M39	0.57	-7.31***
M40	0.64	-8.91***
M41	0.61	-8.12***
M42	0.47	-4.20***
M43	0.62	-8.46***
M44	0.47	-5.35***
***p<0.001 *p<0.05		

28 maddelik testin güvenirliği incelendiğinde Cronbach alfa değeri 0.85 olarak belirlenmiştir Bu durum testin yüksek güvenirlikte ölçme yaptığını göstermektedir. Güvenirlik kat sayısı ayrıca her faktör için test edilmiştir. 1. Faktörün Cronbach Alfa değeri 0.80, 2. Faktörün Cronbach Alfa değeri 0.89 ve 3. Faktörün Cronbach Alfa değeri 0.70 ve 4. Faktörün Cronbach Alfa değeri 0.83 olarak belirlenmiştir. Cronbach Alfa değerleri incelendiğinde alt boyutlarının yüksek güvenirliğe (iç tutarlılık) sahip olduğu gözlenmiştir. Üçüncü faktörün alfa değeri diğer faktörlere göre daha düşüktür. Bu faktörün alfa değeri kabul edilebilir düzeydedir.

Temel bileşenler analizi sonucunda 28 maddeden oluşan ölçek dört faktörden altında toplanmaktadır. 1. Faktör altında 9 madde, 2. Faktör altında 8 madde, 3. Faktör altında 6 madde ve 4. Faktör altında 5 madde yer almaktadır. Birinci faktör "*Öğrenci gelişimini izleme yeterliği"*, ikinci faktör *Öğretimi yönetme yeterliği"*, üçüncü faktör "*Zorlukların üstesinden gelme yeterliği"*, dördüncü faktör ise "*DPA oluşturma yeterliği"* şeklinde adlandırılmıştır.

3.3. DPA öz yeterlik ölçeği doğrulayıcı faktör analizi (DFA)

Araştırmanın üçüncü adımında temel bileşenler analizi yardımı ile dört boyuta indirgediğimiz maddelerin yapı geçerliği DFA analiziyle test edilmiştir. Ölçüm modeli sonuçlarına dayalı olarak aşağıdaki Tablo 5 oluşturulmuştur.

Tablo 5. Ölçüm modeli sonuçları

Faktör/Madde	Standartlaştırılmış Yükler	t-değeri	R^2
Faktör1			
M2	0.52	10.61	0.27
M7	**0.47**	**9.54**	**0.22**
M8	**0.72**	**16.04**	**0.52**
M9	0.65	14.04	0.42
M10	0.57	11.97	0.32
M13	0.53	10.93	0.28
M19	0.70	15.47	0.49
M20	0.69	15.24	0.48
M27	0.58	12.25	0.34
Faktör2			
M29	0.58	12.15	0.34
M32	**0.36**	**7.07**	**0.13**
M33	0.56	11.45	0.31
M39	0.70	15.38	0.49
M40	0.67	14.34	0.45
M41	**0.69**	**14.89**	**0.48**
M43	0.63	13.46	0.40
M44	0.61	13.93	0.37
Faktör3			
M6	0.58	12.86	0.34
M11	**0.65**	**12.98**	**0.42**
M12	0.58	11.51	0.34
M24	0.47	8.92	0.22
M26	0.57	11.25	0.32
M37	**0.38**	**7.08**	**0.14**
Faktör4			
M28	0.71	15.47	0.50
M30	**0.79**	**17.89**	**0.62**
M31	0.75	16.73	0.56
M35	0.55	11.34	0.30
M42	**0.52**	**10.49**	**0.27**

Tablo 5'te standartlaştırılmış yükler her bir gözlenen değişken ile ilgili olduğu gizil değişken arasındaki korelasyonlar hakkında bilgi vermektedir. Faktör1 ile en yüksek korelasyonu M8 (0.72), en düşük korelasyonu M7 (0.47) göstermektedir. Faktör1'de değişkenliğin en çok M8 (R^2=0.52) değişkeni tarafından açıklandığı görülmektedir. Faktör2 ile en yüksek korelasyon gösteren M41 ve en düşük korelasyon gösteren M32 değişkenidir. Bu nedenle R^2 katsayısı en yüksek olan M41 (0.48) değişkenidir. Faktör 3 incelendiğinde M11 değişkeninin Faktör 3 ile 0.65 düzeyinde en yüksek korelasyonu gösterdiği. M37'in ise en düşük korelasyon katsayısına (0.38) sahip olduğu belirlenmiştir. Faktör 3'teki değişkenliğin çoğu (R^2=0.42) M11 tarafından açıklanmaktadır. Faktör 4 incelendiğinde M30 en çok (0.79) ve M42 (0.52) en az korelasyon katsayısına sahiptir. Bu faktörde en fazla değişim M30 (R^2= 0.62) değişkeni tarafından açıklanmaktadır. Ölçüm modelinde gözlenen değişkenlerin hata varyansları incelendiğinde hata varyanslarının 0.37 ile 0.87 arasında değiştiği gözlenmiştir.

Bununla birlikte gözlenen t değerlerinin bütün değişkenler için manidar ($p<0.00$) olduğu hesaplanmıştır. Ölçüm modelinin ilk incelemesinden elde edilen bulgulara göre model uyumunu bozan bir durum yoktur. Ölçüm modelinden elde edilen model uyum indeksleri Tablo 6'da verilmiştir.

Tablo 6. Ölçüm modeli için uyum indeksleri

Uyum ölçüsü	Ölçüt	Değeri	Uyum
X^2 /sd	≤ 3	811.04 / 344 =2.36	İyi
RMSEA	0.05 ≤RMSEA≤0.08	0.057	İyi
SRMR	0.05 ≤SRMR≤0.10	0.06	İyi
NFI	≤0.90	0.91	İyi
NNFI	≤0.90	0.94	İyi
CFI	≤0.90	0.94	İyi
GFI	≤0.90	0.88	Zayıf
AGFI	≤0.90	0.86	Zayıf
CN	≥ 200.00	216.29	Yeterli örneklem

Gözlenen ve beklenen matrisler arasındaki farkı ortaya koyan X^2 değerinin anlamlılığı incelendiğinde bu değerin anlamlı ($p<0.00$) olduğu gözlenmiştir. Bu durum araştırmada ele alınan örneklem büyüklüğünden kaynaklı olabilir. Ancak modelde X^2 /sd değerinin istenilen düzeyde olduğu görülmektedir. Bu nedenle model uyum indekslerinin incelenmesine devam edilmiştir. Model uyum indeksleri bakımından GFI ve AGFI dışındaki diğer indeksler iyi uyum değerine sahipken GFI değeri (0.88) ve AGFI değeri (0.86) zayıf bir uyum göstermektedir.

Model uyumlarına ek olarak analiz çıktılarının önerdiği modifikasyonlar da incelenmiştir. Modifikasyonlarda Faktör1'de yer alan M19'un Faktör 3 ve Faktör 4 ile eşleştirilmesi ve M37'in Faktör2 ve Faktör4 ile eşleştirilmesi önerilmektedir. Sırasıyla yapılan düzenlemelerde ilk olarak X^2 değerini en fazla düşüren M37 ve daha sonra M19 maddesinin modelden çıkarılma durumu incelenmiştir. Elde edilen bulgulara göre sadece X^2 /sd değerinde belli bir iyileşme gözlenirken diğer uyum iyiliği indeklerinde istenilen iyileşme gözlenmemiş, GFI değeri istenilen uyum düzeyine ulaşmamıştır. Bu nedenle araştırmacılar tarafından iki maddenin de modelde kalmasına karar verilmiştir. DFA sırasında herhangi bir modifikasyon yapılmamıştır. Temel bileşenler analizi ile test edilen dört faktörlü yapı doğrulanmıştır.

4. TARTIŞMA VE SONUÇ

Birçok ülke eğitim-öğretim programlarında yapılandırmacı yaklaşımı temel almakta, kağıt kalem testleri içerisinde yer alan açık uçlu, kısa yanıtlı ve çoktan seçmeli sorular gibi ürün temelli geleneksel ölçme-değerlendirme araçlarının yanında portfolyo, proje uygulamaları, performans görevi, kavram haritası gibi süreç temelli (durum belirleme odaklı) tamamlayıcı ölçme-değerlendirme araçlarının kullanımını önemli görmektedir. Bu tamamlayıcı ölçme-değerlendirme araçlarının etkili ve verimli bir şekilde kullanılabilmesi için öğretmenlerin bu konudaki bilgilerinin ve yeterliliklerinin tam olması gerekmektedir. Bilgi ve yeterliklerin belirlenmesi için de bu özellikleri ölçen geçerlik ve güvenirlik kanıtları ortaya konmuş ölçme araçlarına ihtiyaç vardır. Bu sayede daha objektif sonuçlara ulaşılabilir. Bu araştırma kapsamında öğretmenlerin DPA'ya yönelik öz yeterlikler inançlarını ortaya koyan bir ölçek geliştirilmiştir.

Geçerlik ve güvenirlik analizleri dikkate alınarak geliştirilen araçta ilk olarak kapsam geçerliği uzman görüşlerine dayalı olarak sağlanmıştır. Bir ölçeğin istenilen özelliğe dayalı ölçüm yapıp yapmadığı hakkında bilgi veren geçerlik analizlerinin başında kapsam geçerliği gelmektedir (Cronbach. 1970). Ölçeğin yapısını ortaya koymak amacıyla temel bileşenler

analizi yapılmış ve öz yeterlik yapısının dört faktör altında toplandığı, bu faktörlerin değişkenin yaklaşık yüzde ellisini açıkladığı belirlenmiştir. Çok faktörlü desenlerde açıklanan varyans oranının %40 ve %60 olması beklenir (Scherer, Wiebe, Luther ve Adam. 1998; Akt: Tavşancıl. 2005). Elde edilen bulgu belirtilen kaynakları destekler niteliktedir. Ayrıca çizgi grafiğindeki eğimler incelendiğinde dört faktörlü yapı açıkça görülmektedir.

Geliştirilen ölçeğin iç tutarlık kat sayısı hakkında bilgi veren Cronbach alfa değerleri hem ölçeğin kendisi hem de alt boyutlar için yüksek güvenirlik düzeyine sahiptir. Bu durum ölçekte yer alan maddelerin birbirleri ile yüksek ilişki gösterdiğini belirtmektedir. Cronbach (1970) ölçekte yer alan alfa düzeyinin 0.70 üstünde bulunması durumunda ölçeğin yüksek iç tutarlığa sahip olduğunu belirtmektedir.

Maddelerin ayırt edicilik düzeylerinin belirlenmesi amacıyla yapılan alt-üst %27'lik gruplara ayırma yöntemine göre de sadece iki maddenin tutumları olumlu olan ve olumsuz olanları ayıramadığı, düşük korelasyon kat sayısı ile belirlenmiş ve ölçeğe alınmamıştır. Ayırt edicilik değerinin 0.19'dan düşük olması durumunda maddeler gözden geçirilmeli ve eğer düzenlenemiyor ise ölçeğe alınmaması önerilmektedir (Kelley, 1939). Maddelerin geçerliği madde toplam test korelasyonu ile de incelenmiştir. Düşük korelasyon değeri maddenin ölçekten çıkarılmasına işaret etmektedir (Cureton, 1966; Guilford,1953). Bu bulguya göre de aynı iki madde ölçekten çıkarılmıştır.

Ölçeğin yapı geçerliği incelenmiş ve temel bileşenler analizinde ortaya konan yapının tekrar sağlandığı gözlenmiştir. Analiz sonuçları incelendiğinde ki kare değerinin anlamlı çıktığı gözlenmiştir. Ki kare değeri örneklemden etkilenilen bir istatistik olduğundan (Byrne, 2003) model uyum indekslerine bakılmıştır. GFI ve AGFI dışındaki diğer indeksler iyi uyum değerine sahipken GFI değeri ve AGFI değeri zayıf bir uyum göstermektedir. Literatür incelendiğinde AGFI değerinin 0.85'in üstünde olması durumunda "kabul edilebilir uyum" olarak değerlendirilebilir (Raykov ve Marcoulides, 2006; Schermelleh-Engel, Moosbrugger ve Müller, 2003; Vieira, 2011).

DPA'ya yönelik öz yeterlik ölçeği öğretmenlerin DPA'lar hakkındaki öz yeterlik algılarını ortaya koymaktadır. Ölçekten alınan artan puanlar öz yeterlik algılarının yüksekliği, azalan puanlar ise düşüklüğü şeklinde yorumlanabilir. Geliştirilen ölçek, ilgili araştırmalar ve kurumlar tarafından öğretmenlerin öz yeterlik düzeylerini belirlemek ve buna dayalı eğitim ve öğretim etkinliklerine yön vermek amacıyla kullanılabilir. Özellikle tamamlayıcı ölçme ve değerlendirme yöntemleri üzerinde çalışan araştırmacılar öğretmenin DPA'ya yönelik öz yeterliği ve öğrenci performansı arasındaki ilişkileri farklı değişkenleri de dikkate olarak çok boyutlu modellerle inceleyebilir.

Teşekkür

Bu çalışma Aksaray Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi tarafından desteklenmiştir. Proje Numarası: 2015-096

5. KAYNAKLAR

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching*. 53(1), 27-30

Andrade, H. G., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation*, 10(3). Retrieved from http://PAREonline.net/getvn.asp?v=10&n=3

Andrade, H., Wang, X., Du, Y., & Akawi, R. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *The Journal of Educational Research*, 102(4), 287-302.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, 84 (2), 191-215.

Bandura, A. (1994*). Self-efficacy*. In V.S. Ramachaudran (Ed.), Encyclopedia of Human Behavior, (4), 71-81. New York: Academic Press.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading.* Alexandria, VA: ASCD.

Byrne, B. M. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, 34(2), 155-175.

Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: a study at the school level. *Journal of School Psychology*, 44, 473-490.

Cronbach, L. J. (1970). *Essentials of psychological testing.* 3rd Edition , Harper & Row.

Cureton, E. E. (1966). Corrected item-test correlations. *Psychometrik*a, 31, 93-96.

Elias, S., & Loomis, R. (2002). Utilizing need for cognition and perceived self-efficacy to predict academic performance. *Journal of Applied Social Psychology*, 32(8), 1687- 1702.

Guilford, J. P. (1953). The correlation of an item with a composite of the remaining items in a test. *Educational and Psychological Measurement*, 13, 87-93

Jerusalem, M. (2002). Theroretischer Teil - Einleitung I, *Zeitschrift für Pädagogik*, 44, 8-12

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education*, 39 (7), 840-852. doi:10.1080/02602938.2013.875117

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30(1), 17-24.

Likert, R. (1932). A Techniques for the measurement of attitudes. *Archives of Psychology*, 140, 5-53

Metin, M. (2013). Öğretmenlerin performans görevlerini hazırlarken ve uygularken karşılaştığı sorunlar. *Kuram ve Uygulamada Eğitim Bilimleri*, 13(3), 1645- 1673.

Metin, M., & Özmen, H. (2010). Fen ve teknoloji öğretmenlerinin performans değerlendirmeye yönelik hizmet içi eğitim (HİE) ihtiyaçlarının belirlenmesi. *Kastamonu Eğitim Dergisi*, 18(3), 819-838.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation*, 7(3),1-5

Özkan, Ö., Tekkaya, C., & Çakıroğlu, J. (2002). Fen bilgisi aday öğretmenlerin fen kavramlarının anlama düzeyleri, fen öğretimine yönelik tutum ve öz-yeterlik inançları. V. Ulusal Fen Bilimleri Eğitimi Kongresi, ODTÜ, Ankara.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144. doi:10.1016/j.edurev.2013.01.002.

Panadero, E., Jonsson, A., & Strijbos, J. W. (2016). *Scaffolding self-regulated learning through selfassessment and peer assessment: Guidelines for classroom implementation*. In D. Laveault & L. Allal (Eds.), Assessment for learning: Meeting the challenge of implementation (pp. 311–326). Cham: Springer International Publishing.

Popham, W. J. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership*, 55(2), 72-75.

Popham, W. J. (2007). *Classroom Assessment: What Teachers Need to Know*. Pearson Education, 5th Edition, USA.

Raykov T & Marcoulides G. A. (2006). *Fundamentals of structural equation modeling. A first course in structural equation modeling*. 2nd ed. London: Lawrence Erlbaum Associates; p.1-3, 41-3

Reddy, Y., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education,* 35, 435- 448. doi:10.1080/02602930902862859.

Reynolds, J., Smith, R., Moskovitz, C., & Sayle, A. (2009). BioTAP: A Systematic Approach to Teaching Scientific Writing and Evaluating Undergraduate Theses. *BioScience,* 59 (10), 896–903. doi:10.1025/bio.2009.59.10.11

Riggs, I. M. ve Enochs L. G. (1990). Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education,* 74(6), 625-637.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74.

Schwarzer R. (1993). General perceived self-efficacy in 14 cultures. Retrieved on 5-June 2007, at URL http://Web.Fu-Berlin.De/Gesund/Publicat/Ehpscd/Health/World14.Htm

Tabachnick, B. G., & Fidel, L. S. (2001). *Using multivariate statistics*. Fourth Edition. Needham Heights, MA: Allyn & Bacon.

Tavşancıl, E. (2005). *Tutumların ölçülmesi ve SPSS ile veri analizi*. Nobel Yayınları, Ankara

Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, 68, 202-248.

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing and elusive construct. *Teaching and Teacher Education,* 17, 783-805.

Venning, J., & F. Buisman-Pijlman (2013). Integrating assessment matrices in feedback loops to promote research skill development in postgraduate research projects. *Assessment and Evaluation in Higher Education,* 38 (5): 567–579.

Vieira, A. L. (2011). *Preparation of the analysis. Interactive LISREL in practice*. 1st ed. London: Springer.

Wiggins, G. (1991). Standart, not standardization: Evoking quality student work. *Educational Leadership.* 48(5), 18-25.

Yılmaz, M., Köseoğlu, P., Gerçek, C., Soran, H. (2004). Yabancı dilde hazırlanan bir öğretmen öz-yeterlik ölçeğinin Türkçeye uyarlanması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 27, 260-267.

Zimmerman, B.J. (2000). Self-Efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25, 82–91. doi:10.1006/ceps.1999.1016

Tablo Ek-1. Rubriklere yönelik öz yeterlik ölçeği

	Rubrik Özyeterlik Ölçeği	**Kesinlikle katılmıyorum**	**Katılmıyorum**	**Kararsızım**	**Katılıyorum**	**Kesinlikle katılıyorum**
1	DPA'ları sınıf içerisinde nasıl uygulayacağım konusunda yeterli bilgiye sahip olduğuma inanıyorum					
2	Öğretmenlik deneyimime rağmen. DPA hazırlarken zorlanabilirim					
3	Öğrencilerin DPA'ya karşı olumsuz tutumları olsa bile. DPA uygulamalarını kolayca yapabileceğime inanıyorum					
4	DPA'lar ile öğrencilerin performanslarını kolayca değerlendirebilirim.					
5	DPA hazırlama konusunda yeterli bilgiye sahip olduğuma inanıyorum					
6	Öğrenciler daha önce kullanmamış olsalar bile. DPA'lar ile uygulamalar yapabilirim.					
7	Teorik bilgim olmasına rağmen. DPA'ları hazırlarken zorlanabilirim.					
8	Çok zaman aldığından dolayı DPA uygulamaları yaparken zorlanabilirim.					
9	DPA'yı hazırlarken bir sorunla karşılaşırsam üstesinden gelebilirim.					
10	Etkili DPA'lar hazırlayarak öğrenci başarısını artırabilirim.					
11	DPA'ları grup çalışmalarında etkili şekilde kullanabilirim.					
12	DPA'ların kullanım amacını öğrencilere açıklamada zorlanabilirim.					
13	Farklı kaynaklardan bulduğum bir DPA'yı konuya adapte etmekte zorlanabilirim.					
14	DPA uygulamaları ile öğrencilerin konuya yönelik ilgilerini artırabilirim.					
15	Farklı DPA türlerine göre puanlama yapmakta zorlanabilirim.					
16	DPA'lar ile öğrencilerin öğrenme sürecindeki eksikliklerini kolaylıkla belirleyebilirim.					
17	Konuya uygun DPA türünü belirlemede zorlanabilirim.					
18	DPA'yı hazırlarken öğrenme kazanımlarını belirlemekte zorlanabilirim.					
19	Sınıf yönetiminde zorlansam bile. DPA uygulamaları yapabilirim.					
20	Yazılı sınavları DPA ile adil şekilde değerlendirebilirim.					
21	DPA'yı hazırlarken ölçülecek davranışlara karar vermekte zorlanabilirim.					
22	Kalabalık sınıflarda DPA uygulamaları yaparken zorlanabilirim.					
23	Farklı kaynaklardan yararlanarak DPA hazırlama konusunda kendimi geliştirebileceğime inanıyorum.					
24	Hazırladığım DPA'ları öğrencilerin kolayca anlayabileceğine inanıyorum.					
25	DPA uygulamaları ile öğrencilerin konuyu öğrenmeye yönelik endişelerini azaltabileceğime inanıyorum:					
26	Uğraşsam bile. DPA'ları hazırlarken öğrenci davranışlarını ölçecek detaylı tanımlar yapmakta zorlanabilirim.					
27	Öğrencilerin nitelikli çalışmalar ortaya çıkarmalarını sağlayacak DPA'lar hazırlayabilirim.					
28	Eğer uğraşırsam. öğrencilere DPA kullanma becerisini kazandırabilirim.					