ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 327-347

Examination of the Reliability of the Measurements Regarding the Written Expression Skills According to Different Test Theories *

Merve YILDIRIM SEHERYELİ ** Şeref TAN ***

Abstract
The aim of the study is to examine the reliability estimations of written expression skills analytical rubric based on the Classical Test Theory (CTT), Generalizability Theory (GT) and Item Response Theory (IRT) which differ in their field of study. In this descriptive study, the stories of the 523 students in the study group were scored by seven raters. CTT results showed that Eta coefficient revealed that there was no difference between the scoring of the raters ($\eta = .926$); Cronbach Alpha coefficients were over .88. GT results showed that G and Phi coefficients were over .97. The students' expected differentiation emerged, the difficulty levels of the criteria did not change from one student to another, and the consistency between the scores among raters was excellent. In the Item Response Theory, parameters were estimated according to Samejima's (1969) Graded Response Model and item discrimination differed according to the different raters. According to b parameters, for all the raters; individuals are expected to be at least -2.35, -0.80, 0.41 ability level in order to be scored higher than 0, 1 or 2 categories respectively with .50 probability. Marginal reliability coefficients were quite high (around .93). The Fisher Z' statistic was calculated for the significance of the difference between all reliability estimates. GT revealed more detailed information than CTT in the explanation of error variance sources and determination of reliability; while IRT provided more detailed information than CTT in determining the item-level error estimations and the ability level. There was a significant difference between the estimated parameters of CTT and GT in interrater reliability ($p < .05$); there was no significant difference between the parameters predicted according to CTT and IRT ($p > .05$).

Key Words: Classical test theory, generalizability theory, item response theory, interrater reliability, reliability, rubric.

INTRODUCTION

Nowadays, the aim of education is to educate individuals as producers of knowledge in line with the needs of society. Individuals who produce knowledge are, at the same time, critical thinkers, problem-solvers, and researchers. In this respect, changing education policies require a change in the measurement and evaluation methods as well. These changes increased the use of evaluation materials and studies related to the higher level of thinking skills (Kutlu, Doğan & Karakaya, 2014).

There are many ways that enable individuals to demonstrate their high-level skills. However, the most important means of transforming abstract thoughts into concrete form is writing or writing skills. Writing is defined as thinking on thinking. It also allows individuals to expand their thoughts by organizing information (Karatay, 2015).

* This article was written by Merve YILDIRIM SEHERYELİ based on her M.A. thesis in May 2018 at the Institute of Educational Sciences at Gazi University under the supervision of Şeref TAN. In addition, a part of the study is presented as a paper: The examination of the reliability of written expression skills rubric according to classical test and generalizability theories. III. INES Education and Social Science Congress, 21st April - 01st May 2018, Antalya.

** Res. Assist., Hasan Kalyoncu University, Faculty of Education, Gaziantep-Turkey, yldrm.mrv.7806@gmail.com, ORCID ID: 0000-0002-1106-5358
*** Prof. PhD., Gazi University, Institute of Educational Sciences, Ankara-Turkey, sereftan4@yahoo.com, ORCID ID: 0000-0002-9892-3369

To cite this article:
Yıldırım-Seheryeli, M. & Tan, Ş. (2019). Examination of the reliability of the measurements regarding the written expression skills according to different test theories. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 327-347. doi: 10.21031/epod.559470

Received: 30.04.2019
Accepted: 22.07.2019

While studies are carried out to measure written expression skills in developed countries in detail, a common study is not carried out on determining the deficiencies of students in this field in our country. Moreover, the lack of a common writing approach in the teaching process also makes it difficult to follow the development of the students' written expression skills. Therefore, it is not possible to identify learning deficiencies and provide constructive feedback regarding these deficiencies (Karatay, 2015). Therefore, the present study investigates the evaluation of the storytelling, which is one of the written expression skills.

It is necessary to obtain a valid and reliable measurement as well as the suitability of the criterion to make a correct decision about the students. As the errors involved in the measurement process decrease, the reliability of the measurement process increases, and therefore, the accuracy of the decision we make about the individual trait measured increases (Köse, 2014). Therefore, measurement theories somehow differ depending on the purpose of use, limits, and how to use the results of measurement, just as Classical Test Theory (CTT), Generalizability Theory (GT) and Item Response Theory (IRT) differ from one another.

According to the CTT, the score a person receives from any test is the observed score, and this score indicates the degree of presence of the property measured by the test. In addition, when some assumptions are met, the observed score is estimated by the sum of a person's true score and the error score. In CTT, this error score is only one score, which is the sum of random errors that are caused by the individual, measurement expert, the environment, the rater, etc. In the GT, which is an extension of CTT and variance analysis, these error sources are included in the measurement processes in order to control them. The greatest advantage of the GT is that it can partition the variances into different error sources. While CTT is concerned with the reliability of measurements obtained from a group of individuals; GT is concerned with generalizing measurements beyond the measurements, materials, and raters obtained from a group of individuals. Thus, with a single analysis, a single reliability estimation can be made in CTT, and the data can be interpreted under reliability and generalizability. The results obtained by the generalizability study in GT prepare the basis for decision studies so that the effect of the changes in raters, number of items, etc. on reliability estimations can be determined. The accurate estimation of the population where all observation conditions and sources of variability take place provides a new perspective on the difference between reliability and validity. However, validity and reliability studies in CTT require different analyzes. While CTT gives us the variability from all error sources as a single estimate, GT provides the opportunity to examine the error sources such as students, items, and raters together and if there is a variation between students they are called *measurement object*. Measurement object may change depending on the purpose of the study, and it can be item or rater. The way that variation sources (fixed or random facet) are chosen determines the generalizability of the source. The source of a fixed variable is limited to the measurement situation. Therefore, it will be difficult to comment on the generalization of the measurement results even if the source of error decreases, and the measurement accuracy increases. In addition, a single reliability coefficient can be estimated in CTT when relative and absolute assessments are to be taken, while two different reliability coefficients can be estimated in GT according to the fact that individuals are compared to other individuals or treated free from the group. Different patterns can be used depending on whether a source of variability is observed in all conditions of the other source of variability in GT. It is possible to make estimations for all sources of variability when using the crossed design only (Brennan, 2000; Cardinet, Johnson & Pini, 2010; Gulliksen, 1950; Güler, Kaya-Uyanık & Taşdelen-Teker, 2012; Shavelson & Webb, 1991).

In the IRT whole-test and item-level analyses are performed with the relationship between ability estimations and response patterns. In IRT the degree of the latent trait in individuals can be calculated with ability estimations independent from items and with item parameters independent from the sample (Atılgan, 2005; Baykul, 2010; Erkuş, Sünbül, Ömür-Sünbül, Aşiret & Yormaz, 2017). IRT estimates item-based error using the response patterns given to each item. For reliability and validity of three parameter model, the parameters of a, b, c, and θ are examined, and the marginal reliability coefficients are estimated (Baker, 2016; DeMars, 2016). IRT, which is based on fixed variability source, has no purpose of generalizing differently from GT. The difficulty of providing IRT with the

assumptions of unidimensionality and local independence also makes it difficult to use this theory (Ayala, 2009; Hambleton & Jones, 1993; Hambleton, Swaminathan & Rogers, 1991; Ostini & Nering, 2006).

Purpose of the Study

The purpose of the present study is to compare the reliability estimation methods based on CTT, GT, and IRT by using written expression skill scores which are one of the high-level thinking skills of the students and to provide a theoretical contribution to the field by determining their superiorities and differences, limitations and assumptions.

This study is also important in terms of providing the assumptions for the three theories and revealing the findings and interpretations about the difficulties and solutions that the researchers may face concerning the applicability of these theories.

The literature shows that the studies comparing the two theories are more in number than the studies comparing the three theories (Brennan, 2011; Güler, 2008). In the studies which CTT and GT have compared the reliability in terms of internal consistency scores that were obtained from the scales, Kendall's concordance coefficient for non-parametric tests in the occurrence of more than two measures, and G and Phi coefficients obtained by using crossed design in GT were calculated. In general, the results showed that the GT has more detailed results than CTT, and when the number of items and raters increased, the generalizability and reliability coefficients increased as well. For future studies, it is suggested that different items, raters or designs may be used for the same analyses and that the results may be compared by doing analyses in IRT (Bağcı, 2015; Büyükkıdık, 2012; Deliceoğlu, 2009; Güler, 2011; Öztürk, 2011; Şalgam, 2016; Yelboğa & Tavşancıl, 2010).

In studies that compare CTT with IRT, it is generally aimed to compare the item parameters, and it has been observed that large-scale study groups were used with the tests with two-category items. Although they are generally similar in item parameters, it is concluded that IRT provides more detailed results than CTT; CTT is useful in pass-fail decisions; IRT is superior in item invariance or individualized test. Although there is not much research based on reliability comparison, the a and b parameters have been examined on the basis of the item, and it has been seen that reliability interpretations are made only on the item and test functions (Çelen & Aybek, 2013; Doğan & Tezbaşaran, 2003; Gelbal, 1994; İlhan, 2016; Kan, 2006; Kelecioğlu, 2001; Kim & Feldt, 2010; Koch, 1983; Köse, 2015; Lee, Torre & Park, 2012; Morales, 2009; Nartgün, 2002; Özdemir, 2004; Özer-Özkan, 2012; Sebille et al., 2010; Sünbül, 2011).

In the studies comparing the GT and IRT, many facet Rasch measurement model (MFRMM) is generally used. While GT is used to obtain the group and general information, MFRMM is used to obtain information about the sources of variability of items. Apart from examining the sources of variability, the estimation of the reliability coefficients for IRT was not mentioned (Arsan, 2012; Kim & Wilson, 2009; Ure, 2011).

The theories to be used vary depending on the purpose of the researchers, the measurement tool, the data collection method, the measurements obtained, the distribution of measurements, the sampling, the purpose for which the measurements are used and the limitations of the theories. However, a common point of view is that using at least two theories together yields more reliable results. This study compares the CTT, GT, and IRT in the reliability estimation of the scores obtained from a scale which is scored polytomously in line with the suggestions of the studies in the literature.

METHOD

In this study, the techniques used in estimations of reliability in CTT, GT, and IRT methods will be compared by using the story writing skill rubric. This study is a descriptive study, as it just presents

the results as it is without questioning causality or making comparisons and without the effort of determining the relationship or the difference (Erkuş, 2017).

Study Group

The study group consisted of 523 primary and secondary school students. The data were collected in the spring of 2017. One school was in Karabük and the other was in Gaziantep. The distribution of students across province and class levels is as follows:

Table 1. Distribution of Students in the Study Group Across Province and Class

	3rd Grade	4th Grade	5th Grade	6th Grade	7th Grade	Total
Karabük	50	28	18	36	26	158
Gaziantep	52	58	98	74	83	365
Total	102	86	116	110	109	523

Two teachers from Bursa, three from Karabük, one from Gaziantep and one from Ankara volunteered for scoring the data. Work experience of teachers varies between two and ten years. One of them is Turkish teacher, five of them are elementary school teachers, and the last one is an assessment expert.

Data Collection Instruments and Procedure

In this study, the students were asked to write a story according to the criteria given in the determined subjects. Since this practice was done within the class hour, the students and the teachers were chosen voluntarily. The themes of the forms were unanimously voted by three academicians who work in the fields of Elementary School Teaching, Turkish Education and Curriculum Development in Education. The theme for 3rd grade is *forest*, for 4th grade is *colors*, for 5th grade is *books*, for the 6th grade is *teacher*, and for 7th grade is *discrimination*.

Written stories were scored by seven raters according to the written expression skill (analytical) rubric. Each of the raters is provided with the necessary training on how to use the rubric. Scoring range is 0-3, and the highest score that can be obtained from the rubric for 11 criteria is 33, and the lowest score is 0.

Data Analysis

IBM SPSS 22 was used for Eta correlation and Cronbach's Alpha (α) coefficients for CTT, Edu-G 6.1e were used for G and Phi (ϕ) coefficients for GT, and Multilog 7.03 was used for a, b_1, b_2, b_3 (b: parameters of step functions) parameters, and information functions for IRT analysis. In order to compare the reliability coefficients, t-test was performed for the significance of the difference between the two correlation coefficients using Fisher's Z transformation in Microsoft Office Excel 2016 program. For normality assumptions, graphs in IBM SPSS 22, skewness and kurtosis coefficients in Microsoft Office Excel 2016 program were examined. The principal components analysis in SPSS 22 for the assumptions of unidimensionality and local independence were calculated. For model-data fit, the differences between observed and expected ratios in Multilog 7.03 program were investigated.

RESULTS

The skewness and kurtosis coefficients were calculated with Microsoft Excel 2016 before starting analysis under CTT. The skewness coefficients of all grade levels are between -0.612 and 0.873. The kurtosis coefficients are between -1.491 and 0.735. In this case, it can be said that the distribution of data is not skewed and that the kurtosis is acceptable. The results reveal a normal distribution.

Before moving on to the sub-problems of the research, descriptive statistics of the total scores which were scored by seven raters are given below.

Table 2. Descriptive Statistics of the Total Scores Across Grade Levels which were Scored by Seven Raters

Grade Levels	Raters	Min	Max	Mean	Std. Deviation	Grade Levels	Raters	Min	Max	Mean	Std. Deviation
3	1	2	32	11.51	6.456	6	1	0	33	17.84	9.059
N = 102	2	2	32	11.51	6.565	N = 110	2	0	33	17.60	9.201
	3	1	32	11.80	6.236		3	0	33	17.70	9.095
	4	2	32	11.83	6.456		4	0	33	17.95	8.931
	5	2	33	12.28	6.692		5	0	33	18.16	9.064
	6	5	31	**16.45**	5.538		6	4	32	**21.01**	6.905
	7	4	32	14.11	6.038		7	4	33	19.89	8.275
4	1	1	33	12.70	8.889	7	1	4	33	23.28	7.277
N = 86	2	1	33	12.42	8.982	N = 109	2	4	33	22.89	7.288
	3	1	33	12.19	9.145		3	4	33	22.89	6.915
	4	1	33	12.83	9.131		4	4	33	23.05	7.099
	5	1	33	12.51	9.176		5	4	33	22.96	6.987
	6	1	31	**17.06**	6.886		6	3	33	**23.22**	7.186
	7	3	33	15.04	6.364		7	3	33	21.31	7.267
5	1	1	32	14.26	8.012	Total	1	0	33	16.10	9.018
N = 116	2	1	33	14.24	8.285	N = 523	2	0	33	15.92	9.080
	3	1	33	13.90	8.092		3	0	33	15.88	8.948
	4	1	33	14.05	8.325		4	0	33	16.11	9.002
	5	1	33	13.92	8.305		5	0	33	16.15	9.032
	6	6	31	**19.80**	6.264		6	1	33	**19.66**	7.007
	7	4	32	18.37	7.198		7	3	33	17.92	7.601

Table 2 shows that the scores given to the students range between 1.00 and 33.00 in the 3rd, 4th, 5th and 7th-grade levels; however, for the 6th grade, it is between .00 and 33.00. In all levels, the 6th rater scored with a higher mean than the other raters, and as the grade levels increased, the means of the scores given by each rater increased as expected. The most homogeneous scoring was done by 6th rater for the 3rd, 5th, 6th grades and all students (total); by 7th rater for the 4th grade; by 3rd rater for the 7th grade.

Results of Classical Test Theory

The Eta correlation coefficient was calculated using the random block ANOVA results for the consistency of the scores of the seven raters. As a result, it was observed the degree of agreement between the raters who scored the story writing skill of each student was $\eta = .926$ for 11 items, which shows us that the fit among the raters is high. However, this correlation coefficient does not provide us whether each rater scored correctly. For this reason, Cronbach's Alpha (α) internal consistency coefficient was calculated for the reliability of the scores given by the seven raters to the writings of 523 students.

Table 3. Cronbach α Internal Consistency Coefficients for Each Rater

Raters						
1	2	3	4	5	6	7
.936	.936	.934	.937	.938	.880	.901

Table 3 reveals that the scores of each rater are quite high (over .88). In particular, the reliability coefficients of the scores of the first five raters and the seventh rater are considerably high (.90 and above).

Results of Generalizability Theory

In order to calculate the variance and percentages obtained by the G study, seven raters (p) were asked to rate the writings of 523 students (b) using 11 criteria (o), and a completely crossed pattern (b×o×p) was applied. The main effects of b, o and p in this pattern and the effects of bo, bp, op, bop are presented in the table below.

Table 4. Estimated Variances in G study and their Percentages in Total Variance

Source of Variance	df	Sum of Squares	Mean Square	Variance	Percentage
b	522	21401.956	41.000	.525	44.5
o	10	1028.432	102.843	.015	1.3
p	6	72.255	12.043	-.006	.0
bo	5220	4715.283	.903	.059	5.0
bp	3132	565.278	.181	-.028	.0
op	60	2808.594	46.810	.089	7.5
bop	31320	15447.874	.493	.493	41.8
Total					100

It was found that the variance value (.525) estimated for the main effect of the student variable (b) explained 44.5% of the total variance. This variance component for the population score shows how the students differ from each other in a systematic way. The highest (the first rank) value of the variance component is the desired outcome.

The percentage of the estimated error variance (.015) for the main effect of the criterion variable (o) is 1.3%. The low value indicates that there is not much variation among item difficulties.

The percentage of total variance estimation of the predicted error variance value for the main effect of the rater (p) was 0% (-.006, negative values are rounded to zero since the variance cannot be negative). This value gives the degree of variation among the scores of the raters. Because this value is zero, it is an indication of the excellent consistency between the scores of the raters.

The error variance component resulting from the student-criterion (bo) interaction is the difference in students' responses from one criterion to another. The estimated variance value (.059) for this interaction accounted for 5% of the total variance. Accordingly, the difficulty levels of the criteria do not differ much from one student to another.

The error variance component (-.028) resulting from the interaction of the student-rater (bp) explains 0% of the total variance. This value indicates that if a rater gave a high score to a student, other raters gave a high score to that student as well.

The error variance value (.089) resulting from the criterion-scoring (op) interaction accounts for 7.5% of the total variance. This value implies the extent to which a rater is strict when scoring a criterion and flexible when scoring another.

The student-criterion-rater (bop) (residual) variability source is the variability caused by the interaction of the student, the criterion, and the raters and by the random errors. This error variance value (.493), which is the second highest, accounts for 41.8% of the total variance. This value is an indicator of the existence of systematic or random variability sources that cannot be measured in this study by the interaction between students, criteria, and raters.

G and Phi coefficients which are estimated as a result of the decision studies performed by doubling the number of criteria and decreasing it by 2, 6; decreasing the number of raters by 2, 4, 5 and increasing it by 1 are given in the table below.

Table 5. G and Phi (ϕ) Coefficients Obtained from the D Study on Measurement of Written Expression Skills of Students

	Number of the Raters									
	2		3		5		7		8	
Number of criteria	G	Phi	G	Phi	G	Phi	G	Phi	G	Phi
5	.896	.878	.922	.907	.943	.932	.953	.944	.956	.947
9	.939	.928	.955	.946	.968	.961	.973	.968	.975	.970
11	.950	.941	.963	.956	.973	.968	**.978**	**.974**	.980	.975
22	.974	.969	.981	.977	.987	.984	.989	.987	.990	.987

Table 5 shows the result of the real application where 11 criteria were scored by seven raters in which the G coefficient is .978, and ϕ coefficient is .974. The table also reveals that ϕ coefficient is smaller than the G coefficient under similar conditions. Due to the high value of the obtained results, instead of examining the increase in the criteria and raters in D studies, it was tried to obtain values closer to .80 to ensure practicality.

While the smallest G and ϕ coefficients were .896 and .878 respectively when there were five criteria and two raters. The biggest G and ϕ coefficients were .990 and .987, respectively when there were 22 criteria and eight raters. G and ϕ coefficients decreased when the number of raters was decreased, and the number of criteria was fixed. G and ϕ coefficients increased when the number of raters was increased. However, G and ϕ coefficients decreasingly increased after a certain number of items and the raters.

Results of Item Response Theory

One of the polychotomous IRT models: Samejima's graded response model (GRM)

First of all, it is necessary to check the assumptions of IRT. The normality distribution of the data was shown in the CTT analyses. In IRT, the assumptions of unidimensionality and local independence were examined.

In order to check the unidimensionality assumption, Principal Components Analysis (PCA) was performed for each of the seven raters. Eigenvalues, lowest factor loads, and explained variance rates are given in the table below.

Table 6. PCA Results for Unidimensionality Assumption regarding the data of Seven Raters

Rater	The eigenvalue of factor 1	The eigenvalue of factor 2	Proportions of eigenvalues	Assumption of unidimensionality	The lowest factor load	The variance explained by a unidimensional model (%)
1	6.726	1.525	4.41	Provided.	.627	61.144
2	6.726	1.502	4.48	Provided.	.605	61.144
3	6.651	1.588	4.19	Provided.	.615	60.466
4	6.757	1.498	4.51	Provided.	.608	61.426
5	6.792	1.462	4.65	Provided.	.605	61.750
6	5.148	2.293	2.25	Not provided.		
7	5.627	2.028	2.77	Not provided.		

Table 6 indicates that the structure has a dominant dimension for the first five raters since the first eigenvalues are more than four times the second eigenvalues (Çokluk, Şekercioğlu & Büyüköztürk, 2014). Data for 6th and 7th raters could not be included in GRM analysis because they did not meet the assumption of unidimensionality.

If the scale shows the unidimensionality, the assumption of local independence is met as well (Crocker & Algina, 2006), which means that the assumption of local independence is met for the first five raters.

When the observed and expected ratios of each item scored by five raters for model data fit were examined, it was found that the maximum residual value was .0321. Uyar, Öztürk-Gübeş and Kelecioğlu (2013) state that the differences between observed rates and expected rates are named as *residual*. Also, they mention that when the residues approach zero the model – data fit is achieved. Table 7 presents the estimated item parameters and their standard errors according to GRM in measuring the written expression skills.

Table 7. Step-Function Parameters and Standard Errors with the Discrimination of the Items of Written Expressions Rubric

	Raters																			
Items	**1**				**2**				**3**				**4**				**5**			
	a	**b_1**	**b_2**	**b_3**	**a**	**b_1**	**b_2**	**b_3**	**a**	**b_1**	**b_2**	**b_3**	**a**	**b_1**	**b_2**	**b_3**	**a**	**b_1**	**b_2**	**b_3**
1	1.45	-1.15	0.40	1.36	1.40	-0.97	0.29	1.44	1.30	-1.29	0.17	1.45	1.41	-1.15	0.25	1.48	1.52	-1.11	0.29	1.42
SE	0.16	0.14	0.11	0.15	0.15	0.13	0.11	0.16	0.14	0.15	0.11	0.18	0.15	0.14	0.11	0.16	0.16	0.14	0.10	0.14
2	1.54	-1.32	0.06	1.21	1.55	-1.28	-0.02	1.18	1.53	-1.25	-0.05	1.06	1.73	-1.31	0.02	1.05	1.67	-1.23	0.00	1.20
SE	0.16	0.15	0.10	0.13	0.17	0.14	0.10	0.13	0.15	0.13	0.09	0.13	0.17	0.13	0.09	0.12	0.16	0.13	0.09	0.12
3	1.71	-1.33	-0.28	0.68	1.60	-1.42	0.24	0.79	1.74	-1.50	-0.41	0.59	1.88	-1.39	-0.30	0.66	1.88	-1.34	-0.28	0.62
SE	0.18	0.14	0.09	0.11	0.16	0.15	0.09	0.11	0.17	0.14	0.09	0.10	0.18	0.13	0.08	0.09	0.18	0.13	0.08	0.09
4	1.57	-2.17	-0.61	0.44	1.36	-2.04	-0.65	0.52	1.37	-2.35	-0.80	0.41	1.41	-2.22	-0.79	0.45	1.53	-2.15	-0.72	0.46
SE	0.18	0.25	0.10	0.10	0.16	0.24	0.12	0.12	0.16	0.25	0.12	0.11	0.16	0.25	0.12	0.11	0.16	0.23	0.11	0.10
5	1.67	-1.47	-0.14	0.71	1.58	-1.47	-0.20	0.77	1.67	-1.44	-0.35	0.63	1.83	-1.43	-0.31	0.70	1.80	-1.41	-0.27	0.70
SE	0.18	0.16	0.09	0.11	0.17	0.16	0.09	0.11	0.17	0.14	0.09	0.11	0.19	0.14	0.08	0.10	0.18	0.14	0.08	0.10
6	3.05	-0.44	0.44	1.33	2.81	-0.46	0.56	1.51	3.05	-0.56	0.32	1.31	3.17	-0.48	0.35	1.24	2.96	-0.47	0.42	1.28
SE	0.25	0.06	0.06	0.08	0.24	0.07	0.07	0.09	0.25	0.06	0.06	0.09	0.28	0.06	0.06	0.07	0.24	0.07	0.06	0.08
7	4.14	-0.46	0.35	1.11	3.81	-0.47	0.31	1.17	3.74	-0.58	0.20	1.05	4.04	-0.50	0.28	1.04	3.85	-0.50	0.32	1.15
SE	0.34	0.05	0.05	0.06	0.31	0.05	0.06	0.06	0.30	0.05	0.05	0.06	0.31	0.05	0.05	0.06	0.31	0.05	0.05	0.06
8	5.98	-0.44	0.39	0.95	5.89	-0.41	0.34	0.98	5.48	-0.52	0.29	0.94	5.30	-0.43	0.38	0.93	5.52	-0.41	0.36	0.97
SE	0.54	0.04	0.05	0.04	0.50	0.04	0.04	0.04	0.48	0.04	0.05	0.05	0.46	0.04	0.05	0.05	0.49	0.04	0.05	0.04
9	5.73	-0.44	0.32	0.92	6.44	-0.39	0.33	0.97	4.94	-0.50	0.20	0.94	6.20	-0.45	0.22	0.90	6.14	-0.42	0.29	0.93
SE	0.47	0.04	0.05	0.04	0.57	0.04	0.05	0.05	0.44	0.05	0.04	0.05	0.57	0.04	0.04	0.04	0.56	0.04	0.04	0.04
10	5.05	-0.44	0.39	0.95	5.17	-0.46	0.37	0.98	4.81	-0.64	0.26	0.90	4.33	-0.59	0.29	0.95	4.91	-0.54	0.33	0.91
SE	0.42	0.04	0.05	0.05	0.46	0.04	0.05	0.05	0.38	0.04	0.05	0.05	0.35	0.05	0.05	0.05	0.41	0.05	0.05	0.05
11	1.38	-1.74	0.70	1.60	1.26	-1.90	0.72	1.82	1.28	-2.00	0.57	1.87	1.33	-1.88	0.64	1.90	1.24	-1.93	0.74	1.87
SE	0.15	0.22	0.12	0.17	0.15	0.25	0.13	0.19	0.14	0.23	0.12	0.21	0.15	0.23	0.12	0.20	0.15	0.25	0.13	0.20

Table 7 demonstrates that a parameters of the 11 items ranged from 1.24 to 6.44 in all raters. Baker (2016) classified the discrimination parameters as very low (0.01-0.34), low (0.35-0.64), moderate (0.65-1.34), high (1.35-1.69), and very high (1.70 and above). As it is given in Table 7, a parameters in all items are high and very high with regard to the five raters. a parameters show the the slope of item characteristic curve in dichotomous IRT. In polychotomous IRT it additionally shows the item information (DeMars, 2010). According to the 1st and 3rd raters, the most informative item is the 8th one. According to the 2nd, 4th, and 5th raters, the 9th criterion is the most informative one. The least informative item was the 11th criterion according to all the raters.

Table 7 shows the parameters of location related to the step functions (the threshold values for the categories). The b parameters indicate the ability levels of individuals who have been scored into the relevant category by the raters with the probability of .50. Individuals need a lower ability level to be scored into a lower category, while a higher level of skills requires higher categories. For all raters, individuals must have a minimum score of -2.35 ability level to score higher than 0 with .50 probability, and a minimum of -0.80 ability level to score higher than category 1, and a minimum of 0.41 ability level to score higher than category 2.

Characteristic curves and information functions are also used to examine the statistical analysis of items. Item characteristic curves and information functions of the 1st rater are given in appendix A as an example.

Appendix A shows the graphs of items 3 and 8. Since the curves of item 8 are higher (5.98) than the curves of item 3, the discrimination for item 8 is higher; the curves of item 3 are more skewed; therefore, the discrimination for item 3 is lower (1.71).

b parameters of item 8 show that individuals are expected to be at the ability levels of (-∞, about -0.60), [-0.60, about 0.40), [0.40, about 1.20) or [1.20, + ∞) in order to be scored into 0, 1, 2 or 3 categories respectively with .50 probability. Item information function of item 8 revealed that the ability levels in which the item gives the most information are approximately between -0.60 and 1.20.

b parameters of item 3 show that individuals are expected to be at the ability levels of (-∞, about -1.30), [-1.30, about -0.30), [-0.30, about 0.50) or [0.50, + ∞) in order to be scored into 0, 1, 2 or 3 categories respectively with .50 probability. Item information function of item 3 revealed that the ability levels in which the item gives the most information are approximately between -1.50 and 1.00.

In addition to the parameters, the test information function, which is the sum of the contribution of each item to the test, and the marginal reliability coefficient are calculated under IRT. The test information functions of the five raters are given in appendix B.

In appendix B, the figures indicate that even though the test information functions for each rater changes depending on the ability levels, they are relatively higher for individuals with varying ability levels between -1.00 and 1.50. As the amount of information in test information functions increases, the standard error decreases. Then, for individuals who have the ability between -1.00 and 1.50, the measurement results are estimated with fewer errors. As the test information increases, the error level decrease and vice versa.

The marginal reliability coefficient is the coefficient of reliability that is estimated for the whole scale. It takes a value between 0-1; as you get closer to 1, the reliability of the scores obtained from the scale increases. The marginal reliability coefficients of the five raters are given below.

Table 8. Marginal Reliability Coefficients of Five Raters

Raters				
1	2	3	4	5
.9313	.9304	.9313	.9330	.9330

Table 8 shows that all coefficients are around .93 and the reliability is quite high. The Cronbach's Alpha coefficients in CTT for each rater were compared with the marginal reliability coefficients in IRT. In order to compare the Cronbach's Alpha coefficients in the CTT with the coefficients in the GT, the median of the Cronbach Alpha (α) coefficients of the seven scores was calculated.

Table 9. α (median), Eta, G and Phi Coefficients for All Students

α (median)	Eta	G	Phi
.936	.926	.978	.974

The four coefficients in Table 9 are compared in pairs with Fisher's Z test, with .95 probability (.05 significance level). The Z test statistic results for Fisher's Z values and their significance (p) levels are given in the table below:

Table 10. Z Test Results for Fisher's Z Values for Four Coefficients

Fisher Z Coefficients		***α* (median)**	**Eta**	**G**
		1.705	1.35	2.249
Eta	1.35	1.20		
G	2.25	-8.78*	-9.98*	
Phi	2.17	-7.42*	-8.62*	1.36

*p < .05

Table 10 shows that there is no significant difference between the G and Phi coefficients (Z = 1.36, p > .05), and α and Eta correlation coefficients (Z = 1.20, p > .05), while there were significant differences at .05 level between α and G, α and Phi, G and Eta, and finally Phi and Eta correlation coefficients.

According to Table 3 and 8, when the two coefficients were compared with the Z test statistic performed by Fisher's Z conversion, the results obtained with .95 confidence (.05 significance level) are given in the table below:

Table 11. The Results of the Stability Test of the Fisher Z Values of Two Coefficients

Coefficients	**Raters**				
	1	**2**	**3**	**4**	**5**
Fisher Z (α)	1.7047	1.7047	1.6888	1.713	1.721
Fisher Z (Marginal reliability)	1.6681	1.6614	1.6681	1.681	1.681
Z test statistics	.5909	.6996	.3345	.513	.646
p values	.55	.48	.74	.61	.52

Table 11 shows that there is no significant difference between the stability coefficients at .05 level. In this case, the same results were obtained for inter-rater reliability in both CTT and IRT.

DISCUSSION and CONCLUSION

According to the CTT, the Eta correlation coefficient was estimated for the seven raters, and it was seen that the raters' scoring consistency were high. Cronbach α reliability coefficients were high in the internal consistency of the test scores of seven raters. These findings yielded similar results with Cronbach's internal consistency coefficients calculated over .77 in the studies of Bağcı (2015), Büyükkıdık (2012), Deliceoğlu (2009), Güler (2008), Öztürk (2011) and Yelboğa (2007). However, in Güler's (2011) study, the coefficient was very low. Güler (2011) stated that the reason for this result was the purpose of the study and that random data with low validity and reliability were used.

The estimated parameter values in the measurement of written expression skill under GT are explained below.

The error variances and the percentage of total variance estimations that were estimated as a result of the G study of the b×o×p design, in which student (b), criterion (o) and rater (p) variability sources were crossed, were examined.

- It is possible to say that the scoring revealed the variability between the students.
- The criteria do not differ too much from each other as easy, medium, and difficult.
- The consistency between the scoring of the raters is excellent.
- It can be said that the difficulty levels of the criteria do not differ very much from one student to another.
- Students who got high scores from one rater got high scores from others as well.

- Raters can be very strict when scoring a criterion and can be very generous when scoring another. In this study, it was revealed that there are unexplained systematic or random variability sources by design.

These results comply with the results of the studies of Arsan (2012), Brennan (2011), Büyükkıdık (2012), Güler (2008), Deliceoğlu (2009), Şalgam (2016) and Yelboğa (2007). These studies were conducted with a completely crossed design; the number of participants ranged from 72 to 397 and the number of raters ranged from 2 to 9, and data were obtained using likert scales, holistic and analytical rubrics.

G and Phi coefficients obtained as a result of the decision study (D) by increasing and decreasing the number of scoring and criterion in the b×o×p design were examined.

As a result of the real implementation of 11 criteria scored by seven raters, the coefficients G are over .96, and the ϕ coefficients are estimated to be over .95. At the same time, ϕ coefficient was found to be smaller than the G coefficient under similar circumstances as it should be theoretically. Due to the high value of the obtained results, instead of examining the increase in the criteria and raters in D studies, it was tried to obtain values closer to .80 to ensure practicality. These results differ with the studies of Güler (2011) and Öztürk (2011), which had low values of G and Phi coefficients. The reasons for this difference are the fact that Öztürk (2011) used observation form and Güler (2011) used the random data which had low level of reliability and validity.

In this case, GT yields more detailed results than CTT by separating the sources of variability and providing both separate (main) and interactive results including students, criteria, and raters. The literature shows that Çelen and Aybek (2013), Doğan and Tezbaşaran (2003), Gelbal (1994), Kan (2006), Kelecioğlu (2001), Lee et al. (2012), Morales (2009), Nartgün (2002), Özdemir (2004) estimated parameters using dichotomous IRT models with achievement tests or simulated data. Arsan (2012), Ilhan (2016), Kim and Wilson (2009), Özer-Özkan (2012), Sünbül (2011), Ure (2011) estimated parameters using polychotomous-based Rasch model.

According to GRM, a parameters of 11 items ranged from 1.24 to 6.44 in all raters and discriminations of items for each rater and information that items provide are high. According to the 1st and 3rd raters, the item that gives the most information is the 8th criterion, and according to the 2nd, 4th and 5th raters, it is 9th criterion that gives the most information. The least informative item was the 11th criterion in all the raters. Of Koch's (1983), Köse (2015), Nartgün's (2002) and Özdemir's (2004) studies, which are based on polychotomous IRT model, the highest value of item discrimination is 3.34, estimated for the first item of the sample consisted of all males in Nartgün's (2002) study. In this study, the reason for the fact that discrimination value is 6.44 can be because of the academic achievement levels of the schools in the sample, the familiarity of the students to the written expression studies and the inclusion of all students between the 3rd and 7th grades.

Although the difficulty levels of the items do not differ much according to the GT, b parameters according to IRT vary between -2.35 and 1.90 and θ levels vary from -0.50 to 1.20.

It was revealed that marginal reliability coefficients were quite high (around .93). This finding is very close to the marginal reliability coefficients obtained by Köse (2015) and Nartgün (2002) (.97 and .93), whereas it differs from the coefficients (between .65 to .73) obtained by Özdemir (2004). Apart from the marginal reliability coefficient, a single coefficient for reliability in IRT was calculated in Morales (2009) (Person reliability .95) and Çelen and Aybek (2013) (Empirical reliability .80).

Similar to CTT and GT, the reliability of the scores of the five raters in the IRT was high. As a result, the reliability estimates obtained from the three reliability theories used for our measurements were all very high.

The reliability estimates of the three measurement theories used in this study were examined in two ways. There was a significant difference between α and G, α and Phi, G and Eta, Phi and Eta coefficients ($p < .05$) at each grade level and all students in favor of GT. In this case, CTT and GT coefficients differed in reliability estimation. The literature in this field shows that there has been no

analysis for the significance of the difference between the correlation coefficients in the studies comparing CTT to GT.

In the second part, Cronbach Alpha coefficients in CTT and marginal reliability coefficients in IRT were compared. There was no significant difference between the coefficients ($p < .05$). Similar results were obtained for inter-rater reliability in CTT and IRT. Nartgün (2002) examined the difference between Cronbach's Alpha internal consistency coefficient and the marginal reliability coefficient with Fisher Z transformation and found no significant difference. In contrast to this study, Doğan and Tezbaşaran (2003) examined the significance of the difference between item discriminations and difficulties in CTT and IRT with Fisher's Z transformation and concluded that there was no significant difference.

As a result of the present study which aimed to estimate the reliability of the measurements, it was revealed that when the number of samples is at least 500 and the unidimensionality-local independence assumptions are met, making item-level error estimations with Samejima's (1969) Graded Response model and making reliability estimates through item and test information functions in IRT provide more detailed information than those provided by CTT. Unlike CTT, when the number of samples is less than 500 and the variability sources are more than two, it is possible to calculate the generalizability and reliability coefficients, which differ based on the absolute and relative decisions, by examining the error variances separately and together using GT. In studies in which there is a single source of variability, the use of CTT is more useful if there are pass-fail decisions or when the researcher has a purpose of ranking.

REFERENCES

Arsan, N. (2012). *Buz pateninde hakem değerlendirmelerinin genellenebilirlik kuramı ve Rasch modeli ile incelenmesi* (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Atılgan, H. (2005). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama. *Eğitim Bilimleri ve Uygulama, 4*(7), 95-108.

Ayala, R. J. (2009). *The theory and practice of item response theory*. USA: The Guildford Press.

Bağcı, V. (2015). *Matematiksel muhakeme becerisinin ölçülmesinde klasik test kuramı ile genellenebilirlik kuramındaki farklı desenlerin karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Baker, F. B. (2016). *Madde tepki kuramının temelleri* (Çev. N. Güler ve M. İlhan). Ankara: Pegem Akademi.

Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* Ankara: Pegem Akademi.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement And Education, 24*, 1-21. doi: 10.1080/08957347.2011.532417

Brennan, R.L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339–353.

Büyükkıdık, S. (2012). *Problem çözme becerisinin değerlendirilmesinde puanlayıcılar arası güvenirliğin klasik test kuramı ve genellenebilirlik kuramına göre karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. USA: Routledge-Taylor & Francis Group.

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

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik spss ve lisrel uyglamaları.* Ankara: Pegem Akademi.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* USA: Cengage Learning.

Deliceoğlu, G. (2009). *Futbol yetilerine ilişkin dereceleme ölçeğinin genellenebilirlik ve klasik test kuramına dayalı güvenirliklerinin karşılaştırılması* (Yayımlanmış doktora tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

DeMars, C. (2016). *Madde tepki kuramı* (Çev. E. H. Özberk ve H. Kelecioğlu). Ankara: Nobel Akademi.

Doğan, N., & Tezbaşaran, A. A. (2003). Klasik test kuramının ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 25*, 58-67.

Erkuş, A. (2017). *Bilimsel araştırma süreci.* Ankara: Seçkin.

Erkuş, A., Sünbül, Ö., Ömür Sünbül, S., Aşiret, S., & Yormaz, S. (2017). *Psikolojide ölçme ve ölçek geliştirme II.* Ankara: Pegem Akademi.

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

Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma* (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Güler, N. (2011). Rastgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenirliğin karşılaştırılması. *Eğitim ve Bilim, 36*(162), 225-234.

Güler, N., Kaya-Uyanık, G., & Taşdelen-Teker, G. (2012). *Genellenebilirlik kuramı.* Ankara: Pegem Akademi.

Gulliksen, H. (1950). *Theory of mental tests.* USA: John Wiley & Sons.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement Issues and Practice, 12*(3), 38-47. doi: 10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* USA: Sage Publications.

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 346-368. doi: 10.16986/HUJE.2016015182

Kan, A. (2006). Klasik test teorisine ve örtük özellikler teorisine göre kestirilen madde parametrelerinin karşılaştırılması üzerine ampirik bir çalışma. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 2*(2), 227-235.

Karatay, H. (2015). Süreç temelli yazma modelleri: 4+1 Planlı yazma ve değerlendirme modeli. M. Özbay (Ed.), *Yazma eğitimi* içinde (s. 21-48). Ankara: Pegem Akademi.

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

Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review Journal, 11*(2), 179-188. doi: 10.1007/s12564-009-9062-8

Kim, S., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and many-facet rasch measurement. *Journal of Applied Measurement, 10*(4), 408-422.

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7*(1), 15-32. doi: 10.1177/014662168300700104

Köse, A. (2014). Ölçmede güvenirlik. R. N. Demirtaşlı (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (s. 86-109). Ankara: Edge Akademi.

Köse, A. (2015). Aşamalı tepki modeli ve klasik test kuramı altında elde edilen test ve madde parametrelerinin karşılaştırılması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 15*(2), 184-197.

Kutlu, Ö., Doğan, C., & Karakaya, İ. (2014). *Ölçme ve değerlendirme performansa ve portfolyoya dayalı durum belirleme.* Ankara: Pegem Akademi.

Lee, Y.-S., Torre, J. d., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: An empirical investigation. *Asia Pacific Educ. Rev. 13*(2), 333–345. doi: 10.1007/s12564-011-9196-3

Morales, R. A. (2009). Evaluation of mathematics achievement test: A Comparison between CTT and IRT. *The International Journal of Educational and Psychological Assessment, 1*(1), 19-26.

Nartgün, Z. (2002). *Aynı tutumu ölçmeye yönelik likert tipi ölçek ile metrik ölçeğin madde ve ölçek özelliklerinin klasik test kuramı ve örtük özellikler kuramına göre incelenmesi* (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models.* USA: Sage Publication.

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

Özer-Özkan, Y. (2012). *Öğrenci başarılarının belirlenmesi sınavından (ÖBBS) klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması* (Yayımlanmamış doktora tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Öztürk, M. E. (2011). *Voleybol becerileri gözlem formu ile elde edilen puanların genellenebilirlik ve klasik test kuramına göre karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Şalgam, A. (2016). *Kısa cevaplı matematik yazılı sınavının genellenebilirlik kuramı ve test tekrar test yöntemiyle güvenirliğinin kıyaslanması.* (Yayımlanmamış yüksek lisans tezi). Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores [Monograph]. *Psychometrika, 34*(4, Pt. 1). doi: 10.1007/BF03372160

Sebille, V., Hardouin, J.-B., Neel, T. L., Kubis, G., Boyer, F., Guillemin, F., & Falissard, B. (2010). Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients - A simulation study. *BMC Medical Research Methodology, 10*. doi: 10.1186/1471-2288-10-24

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A Primer.* USA: Sage Publications.

Sünbül, Ö. (2011). *Çeşitli boyutluluk özelliklerine sahip yapılarda, madde parametrelerinin değişmezliğinin klasik test teorisi, tek boyutlu madde tepki kuramı ve çok boyutlu madde tepki kuramı çerçevesinde incelenmesi* (Yayımlanmamış doktora tezi). Mersin Üniversitesi Eğitim Bilimleri Enstitüsü, Mersin.

Ure, A. C. (2011). *The effect of raters and rating conditions on the reliability of the missionary teaching assessment* (Unpublished master thesis). University of Brigham Young, USA.

Uyar, Ş., Öztürk-Gübeş, N., & Kelecioğlu, H. (2013). PISA 2009 tutum anketi madde puanlarının aşamalı tepki modeli ile incelenmesi. *Eğitim ve Öğretim Araştırmaları Dergisi, 2*(4), 125-134.

Yelboğa, A. (2007). *Klasik test kuramı ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi* (Yayımlanmış doktora tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Yelboğa, A., & Tavşancıl, E. (2010). Klasik test ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 10*(3), 1825-1854.

Yazılı Anlatım Becerisine İlişkin Ölçümlerin Güvenirliğinin Farklı Test Kuramlarına Göre İncelenmesi

Giriş

Günümüzde eğitimin amacı kişileri, toplumun ihtiyaçları doğrultusunda, tüketen değil bilgiyi üreten bireyler olarak yetiştirmektir. Bilgi üretebilen nitelikteki kişilerin sorunları çözebilen, sorgulayan, üst düzey ve eleştirel düşünebilen, araştırma-geliştirme becerisine sahip ve yaratıcı bireyler olması gerekmektedir. Bireylerin üst düzey becerilerini ortaya koymaları sağlayan birçok araç vardır. Fakat soyut durumdaki düşünceleri somutlaştırarak incelenebilir hâle dönüştüren en önemli araç yazma ya da yazılı anlatım becerisidir. Yazma, düşünme üzerine düşünme olarak tanımlanmaktadır. Ayrıca bireylerin bilgiyi düzenleyerek düşüncelerini genişletmelerini sağlamaktadır (Karatay, 2015).

Gelişmiş ülkelerde yazılı anlatım becerilerinin detaylı olarak ölçülmesi için çalışmalar yapılırken ülkemizde henüz öğrencilerin bu yöndeki eksikliklerinin belirlenmesi üzerine ortak bir çalışma yapılmamaktadır. Ölçmenin yanında öğretim sürecinde de ortak bir yazma yaklaşımının olmaması, öğrencilerinin yazılı anlatım becerilerinin gelişimlerinin takip edilmesini de güçleştirmektedir. (Karatay, 2015). Bu nedenle, bu çalışmada da yazılı anlatım becerilerinden biri olan hikâye yazma becerilerinin değerlendirilmesi konu edinilmiştir.

Öğrenciler hakkında doğru karar verebilmek (değerlendirme yapmak) için ölçütün uygunluğunun yanı sıra geçerli ve güvenilir bir ölçüm de elde etmek gerekmektedir. Ölçme işlemine karışan hatalar azaldıkça ölçme işleminin güvenirliği dolayısıyla da bireyde ölçülen özellik hakkında verdiğimiz kararın doğruluğu artmaktadır (Köse, 2014). Bu nedenle hata teorileri (kuramlar); kullanım amacına, sınırlıklarına, ölçme sonuçlarının ne şekilde kullanılacağına göre Klasik Test Kuramı (KTK), Genellenebilirlik Kuramı (GK) ve Madde Tepki Kuramı (MTK) gibi farklılaşmıştır.

Araştırmacıların, araştırmanın amacına, kullanılan ölçme aracına, veri toplama yöntemine, elde edilen ölçümlere, ölçümlerin dağılımına, örnekleme, ölçümlerin hangi amaçla kullanılacağına, kuramların sınırlıklarına bağlı olarak kullanılması önerilen kuramlar da değişmektedir. Ortak bir bakış açısı, en az iki kuramın birlikte kullanılmasının daha güvenilir sonuçlar ortaya koyduğu yönündedir. Bu araştırma ile öğrencilerin üst düzey düşünme becerilerinden biri olan yazılı anlatım becerisi puanları kullanılarak KTK, GK ve MTK'ye dayalı güvenirlik kestirme yöntemlerinin karşılaştırılması, birbirlerine göre üstünlükleri ve farkları, sınırlıkları ve sayıltıları belirlenerek alana kuramsal bir katkı

sağlanması hedeflenmektedir. Bu çalışma, incelenen üç kuram için sayıltıların sağlanması ve bu kuramların uygulanabilirliğine yönelik olarak araştırmacıların karşılaşabileceği güçlükler ve çözüm yollarına yönelik bulgu ve yorumların yapılması bakımından da önem taşımaktadır.

Yöntem

Araştırmanın çalışma grubunu 2017 yılı bahar döneminde Karabük ve Gaziantep'te bulunan birer okulda öğrenim gören toplam 523 ilkokul ve ortaokul öğrencisi oluşturmaktadır. Bu öğrencilerin 102'si 3. sınıfta, 86'sı 4. sınıft, 116'sı 5. sınıfta, 110'u 6. Sınıfta ve 109'u 7. sınıfta öğrenim görmektedir.

Çalışma grubunda verileri puanlamak için Bursa'dan 2, Karabük'ten 1, Gaziantep'ten 1 ve Ankara'dan 1 kişi olmak üzere toplam 7 öğretmen gönüllü olmuştur. Öğretmenlerin iş tecrübesi 2 ile 10 yıl arasında farklılaşmaktadır. Öğretmenlerimizden biri Türkçe, beşi sınıf öğretmeni ve biri ölçme değerlendirme uzmanı olarak görev yapmaktadır.

Veri toplama araçları

Bu çalışmada öncelikle öğrencilerden belirlenen konularda verilen ölçütlere göre hikâye yazmaları istenmiştir. Bu uygulama ders saati içinde yapıldığından, öğrencilerin ve öğretmenlerin seçilmesinde gönüllülük esas alınmıştır. Formların temaları Sınıf Öğretmenliği, Türkçe Eğitimi ve Eğitimde Program Geliştirme alanlarında çalışmalar yapan üç akademisyen tarafından oy birliği ile 3. sınıf için *orman*, 4. sınıf için *renkler*, 5. sınıf için *kitaplar*, 6. sınıf için *öğretmen*, 7. sınıf için *ayrımcılık* olarak belirlenmiştir.

Yazılan hikâyeler, yazılı anlatım becerisi (analitik) puanlama anahtarına göre yedi puanlayıcı tarafından puanlanmıştır. Puanlayıcıların her birine puanlama anahtarının nasıl kullanılacağı ile ilgili gerekli eğitimler verilmiştir. 0-3 arasında yapılan puanlamada 11 ölçüt için puanlama anahtarından alınabilecek en yüksek puan 33 en düşük puan 0 olarak belirlenmiştir.

Veri analizi

Güvenirlik belirlemede KTK'de Eta korelasyon ve Cronbach Alfa (α) katsayıları için SPSS 22; GK'de G ve Phi (ϕ) katsayıları için Edu-G 6.1e ve MTK'de a, b_1, b_2, b_3 (b: adım fonksiyonlarının parametreleri) ve θ parametreleri ile bilgi fonksiyonları için Multilog 7.03 programları kullanılmıştır. Elde edilen güvenirlik katsayılarının karşılaştırılması için ise Microsoft Ofis Excel 2016 programında Fisher'in Z dönüştürmesi kullanılarak iki korelasyon katsayısı arasındaki farkın manidarlığı için t testi yapılmıştır. Normallik sayıltısı için SPSS 22'de grafikler, Microsoft Ofis Excel 2016 programında çarpıklık ve basıklık katsayıları; tek boyutluluk ve yerel bağımsızlık sayıltıları için yine SPSS 22'de temel bileşenler analizi; model-veri uyumu için ise Multilog 7.03 programında gözlenen ve beklenen oranlar arasındaki farklar incelenmiştir.

Sonuç ve Tartışma

KTK'ye göre, yedi puanlayıcı için puanlayıcılar arasındaki Eta korelasyon katsayısı hesaplanmıştır ve puanlayıcıların öğrencileri puanlamadaki uyumlarının yüksek olduğu görülmüştür. Yedi puanlayıcının da test puanlarının iç tutarlılık olarak Cronbach α güvenirlik katsayıları yüksek bulunmuştur. Bu bulgular Bağcı (2015), Büyükkıdık (2012), Deliceoğlu (2009), Güler (2008), Öztürk (2011) ve Yelboğa' nın (2007) çalışmalarında .77'nin üzerinde hesapladıkları Cronbach α iç tutarlılık katsayıları ile benzer sonuçlar vermiş, farklı olarak Güler'in (2011) rastgele veriler üreterek yaptığı çalışmada çok düşük düzeyde bulunmuştur. Güler (2011) bu durumun sebebinin çalışmanın amacından kaynaklandığını, düşük geçerlik ve güvenirliğe sahip rastgele verilerin kullanıldığını belirtmiştir.

GK'ye göre yazılı anlatım becerisinin ölçülmesinde kestirilen parametre değerleri aşağıda açıklanmıştır.

Öğrenci (b), ölçüt (ö) ve puanlayıcı (p) değişkenlik kaynaklarının tümüyle çaprazlandığı b×o×p deseninin G çalışması sonucunda kestirilen hata varyansları ve toplam varyansı açıklama yüzdeleri incelenmiştir.

- Yapılan puanlamaların öğrenciler arasındaki farklılaşmayı ortaya çıkardığını söylemek mümkündür.
- Ölçütler kolay, orta ve zor gibi birbirinden güçlük bakımından çok fazla farklılaşmamaktadır.
- Puanlayıcıların puanlamaları arasındaki tutarlılık mükemmel düzeydedir.
- Ölçütlerin güçlük düzeylerinin bir öğrenciden diğerine çok büyük farklılıklar göstermediği söylenebilir.
- Bir puanlayıcının yüksek puan verdiği öğrenciler diğer puanlayıcılardan da yüksek puan almıştır.
- Puanlayıcıların bir ölçütü puanlarken çok katı, diğer ölçütte ise cömert olabildikleri görülmektedir. Bu çalışmada ölçülemeyen sistematik ya da tesadüfi değişkenlik kaynaklarının bulunduğu saptanmıştır.

Bu sonuçlar Arsan (2012), Brennan (2011), Büyükkıdık (2012), Güler (2008), Deliceoğlu (2009), Şalgam (2016) ve Yelboğa'nın (2007) tümüyle çaprazlanmış desende 72 ile 397 arasında birey, 2 ile 9 arasında puanlayıcı, likert ölçekler, bütüncül ve analitik rubrik kullanarak elde ettikleri veriler ile örtüşmektedir.

b×o×p deseninde puanlayıcı ve ölçüt sayılarının arttırılıp azaltılmasıyla yapılan karar çalışması (K) sonucunda elde edilen G ve Phi katsayıları incelenmiştir.

11 ölçütün yedi puanlayıcı tarafından puanlandığı asıl uygulama sonucunda G katsayılarının .96'nın üzerinde, ϕ katsayılarının .95'in üzerinde kestirildiği görülmektedir. Aynı zamanda teorik olarak olması gerektiği gibi benzer durumlar altında her ϕ katsayısı, G katsayısından küçük bulunmuştur. Elde edilen sonuçların yüksek değerlerde olması sebebi ile K çalışmalarında ölçüt ve puanlayıcı sayılarının artışlarını incelemek yerine kullanışlılık (ekonomiklik) sağlanması adına daha az puanlayıcı ve ölçüt ile .80'e yakın değerler elde edilmeye çalışılmıştır. Bu sonuçlar ise G ve Phi katsayıları düşük düzeyde elde edilen Güler (2011), Öztürk'ün (2011) çalışmaları ile farklı durumlar ortaya koymuştur. Bu durumun sebebi ise Öztürk'ün (2011) çalışmasında gözlem formu, Güler'in (2011) çalışmasında geçerlik ve güvenirliği düşük olması istenen rastgele veriler olarak belirtilmiştir.

Bu durumda GK, değişkenlik kaynaklarını ayrıştırarak öğrenciler, ölçütler ve puanlayıcıları ayrı ayrı ve etkileşimlerini içeren sonuçlarla KTK'ye göre daha ayrıntılı sonuçlar vermiştir.

Alanyazın incelendiğinde Özdemir (2002), Nartgün (2002), Doğan ve Tezbaşaran (2003), Kan (2006), Morales (2009), Gelbal (1994), Kelecioğlu (2001), Lee, Torre ve Park. (2012), Çelen ve Aybek'in (2013) araştırmalarında başarı testleri ya da simülasyon ile üretilmiş veri kullanarak iki kategorili MTK modelleri; Özer-Özkan (2012), Sünbül'ün (2011) çok boyutlu MTK modelleri; Arsan (2012), İlhan (2016), Kim ve Wilson (2009), Ure'nin (2011) çok değişkenlik kaynaklı Rasch modeli kullanarak parametre kestirimleri yaptıkları görülmüştür.

Derecelendirilmiş (Aşamalı) Tepki Modeli'ne (DTM) göre, 11 maddenin a parametrelerinin tüm puanlayıcılarda 1.24 ile 6.44 arasında değiştiğinden her puanlayıcı için maddelerin ayırt ediciliklerinin ve verdikleri bilgilerin yüksek düzeyde olduğu görülmüştür. 1 ve 3. puanlayıcılara göre en fazla bilgiyi veren madde 8. ölçüt iken 2, 4 ve 5. puanlayıcılara göre en fazla bilgiyi 9. ölçüt vermektedir. En az bilgi veren madde ise tüm puanlayıcılara göre 11. ölçüt olarak bulunmuştur. Çok kategorili MTK modellerinin kullanıldığı Koch (1983), Köse (2015), Nartgün (2002) ve Özdemir (2002)'nin çalışmalarında madde ayırt edicilik değerleri en yüksek Nartgün'ün (2002) çalışmasında erkek örnekleminden birinci madde için kestirilen 3.34 değeridir. Bu çalışmada ise ayırt edicilik değerinin 6.44 bulunması örneklemdeki okulların eğitim düzeyleri, öğrencilerin yazılı anlatım çalışmalarına

aşinalığı, 3 ile 7. sınıflar arasındaki tüm öğrencilerin örnekleme dâhil edilmesinin olabileceği düşünülmektedir.

GK'ye göre maddelerin güçlük düzeyleri çok fazla farklılaşmıyor olarak bulunmasına rağmen MTK'ye göre b parametreleri -2.35 ile 1.90 ve θ düzeyleri -0.50 ile 1.20 arasında farklılaşmaktadır.

Marjinal güvenirlik katsayıları incelendiğinde ise güvenirliğin oldukça yüksek (.93 civarında) olduğu görülmüştür. Bu bulgu Köse (2015) ve Nartgün'ün (2002) elde ettikleri marjinal güvenirlik katsayıları ile çok yakınken (.97 ve .93) Özdemir'in (2002) elde ettiği katsayılardan (.65 ile .73 arasında) farklılaşmaktadır. Marjinal güvenirlik katsayısı dışında MTK'de güvenirlik için tek bir katsayıya Morales (2009) -Person reliability (kişi güvenirliği) .95- ve Çelen ve Aybek'in (2013) -Empirical reliability (Görgül güvenirlik) .80- çalışmalarında rastlanmıştır.

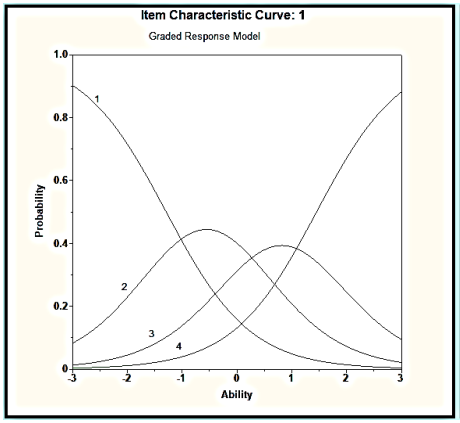
KTK ve GK ile benzer şekilde MTK'de da beş puanlayıcının öğrencilerin yazılı anlatım becerilerini puanlamaları sonucu elde edilen puanların güvenirliği yüksek düzeyde bulunmuştur. Sonuçta ölçümlerimiz için kullanılan üç güvenirlik kuramlarından elde edilen güvenirlik kestirimlerinin hepsi oldukça yüksek bulunmuştur.

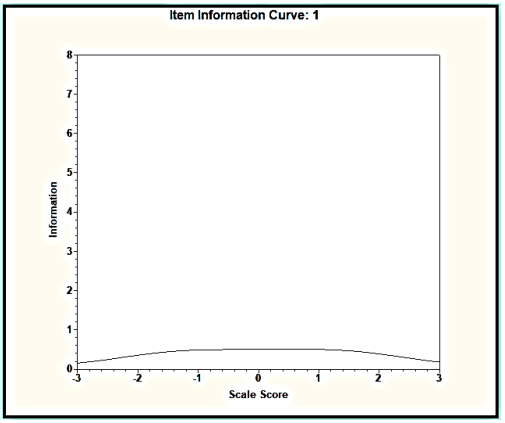
Bu çalışmada kullanılan üç ölçme teorisinden elde edilen güvenirlik kestirimleri arasında manidar bir farklılık olup olmadığı iki şekilde incelenmiştir. İlk kısımda tüm öğrencilere ve tek uyum puanına göre elde edilen KTK'deki Cronbach Alpha katsayıları, Eta korelasyon katsayıları ile GK'deki G ve Phi katsayıları karşılaştırılmıştır. Bu işlem için yedi puanlayıcıya ait Cronbach Alfa (α) katsayılarının ortancası alınmıştır. Her sınıf düzeyinde ve tüm öğrencilerde α ile G, α ile Phi, G ve Eta, Phi ve Eta katsayıları arasında .05 düzeyinde Genellenebilirlik Kuramı katsayıları lehine anlamlı bir fark bulunmuştur. Bu durumda güvenirlik kestirimi için KTK ile GK katsayılarının farklılaştığı görülmüştür. Alanyazın incelendiğinde KTK ile GK'yi karşılaştıran çalışmalarda korelasyon katsayılarının arasındaki farkın manidarlığı için yapılan bir analizle karşılaşılmamıştır.

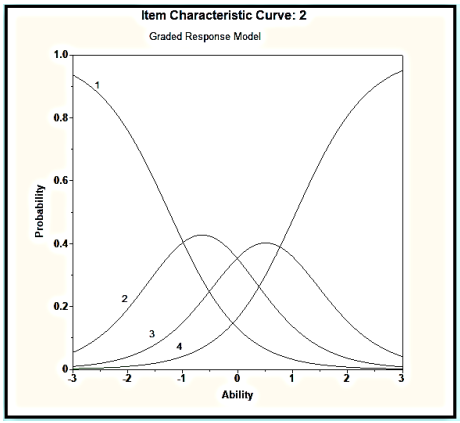
İkinci kısımda ise puanlayıcılara göre elde edilen KTK'deki Cronbach Alpha katsayıları ile MTK'deki marjinal güvenirlik katsayıları karşılaştırılmıştır. .05 anlamlılık düzeyinde katsayılar arasında anlamlı bir fark olmadığı tespit edilmiştir. Bu durumda KTK ve MTK'ye göre puanlayıcılar arası güvenirlik için benzer sonuçlar elde edilmiştir. Alanyazın incelendiğinde KTK ile MTK'yi karşılaştıran çalışmalarda Nartgün (2002) bu çalışma ile benzer olarak Cronbach Alfa iç tutarlılık katsayısı ile marjinal güvenirlik katsayısı arasındaki farkı Fisher'in Z dönüşümü ile inceleyerek manidar bir fark olmadığı sonucuna ulaşmıştır. Doğan ve Tezbaşaran (2003) ise bu çalışmadan farklı olarak KTK ve MTK'deki madde ayırt edicilikleri ve güçlükleri arasındaki farkın manidarlığını Fisher'in Z dönüşümü ile incelemiş, manidar bir fark olmadığı sonucuna ulaşmıştır.

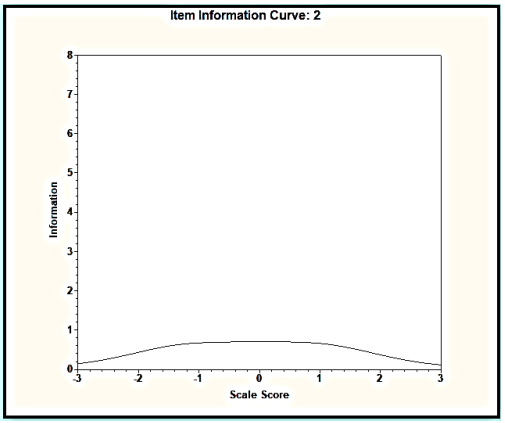
Sonuç olarak ölçümlerin güvenirliğini kestirmeye yönelik olan bu çalışmaya göre, örneklem sayısı en az 500 olduğunda ve tek boyutluluk-yerel bağımsızlık varsayımları karşılandığında MTK'de Samejima'nın (1969) Derecelendirilmiş Tepki modeli ile madde düzeyinde hata kestirimleri yapmak madde ve test bilgi fonksiyonları aracılığıyla güvenirlik kestirimleri yapmak KTK'ye göre daha ayrıntılı bilgiler sunmaktadır. Örneklem sayısı 500'den az, değişkenlik kaynakları ikiden fazla olduğunda, GK kullanılarak hata varyanslarının ayrı ayrı ve birlikte ele alınması ile mutlak ve bağıl kararlara göre farklılaşan genellenebilirlik ve güvenirlik katsayıları hesaplamak KTK'den farklı olarak mümkün olmaktadır. Değişkenlik kaynağının tek olduğu çalışmalarda, geçti-kaldı kararları ya da araştırmacının sıralama yapma amacı olduğunda ise KTK'nin kullanılması daha kullanışlıdır.

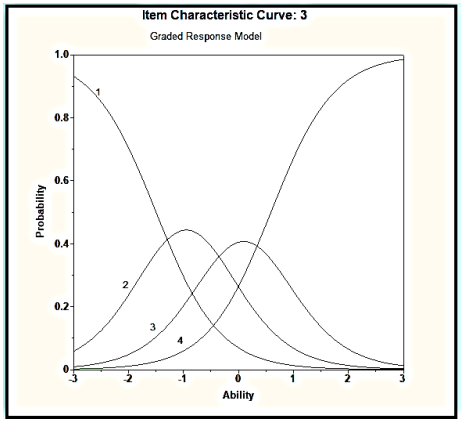
Appendix A. Characteristic Curves and Information Functions of Items-1st Rater

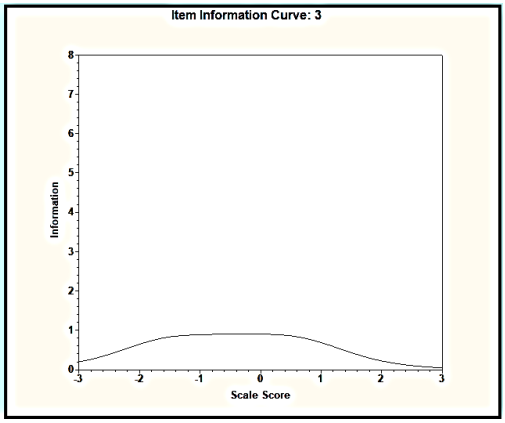


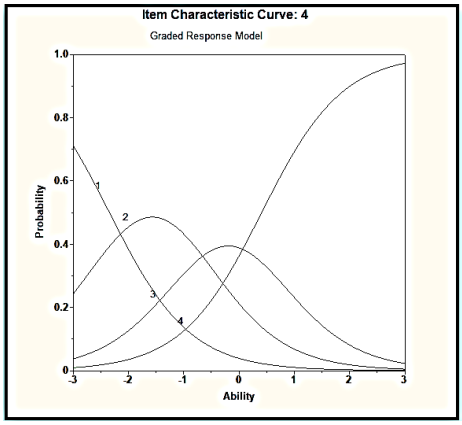


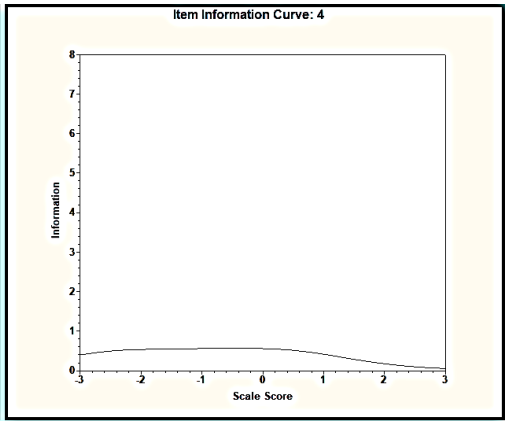


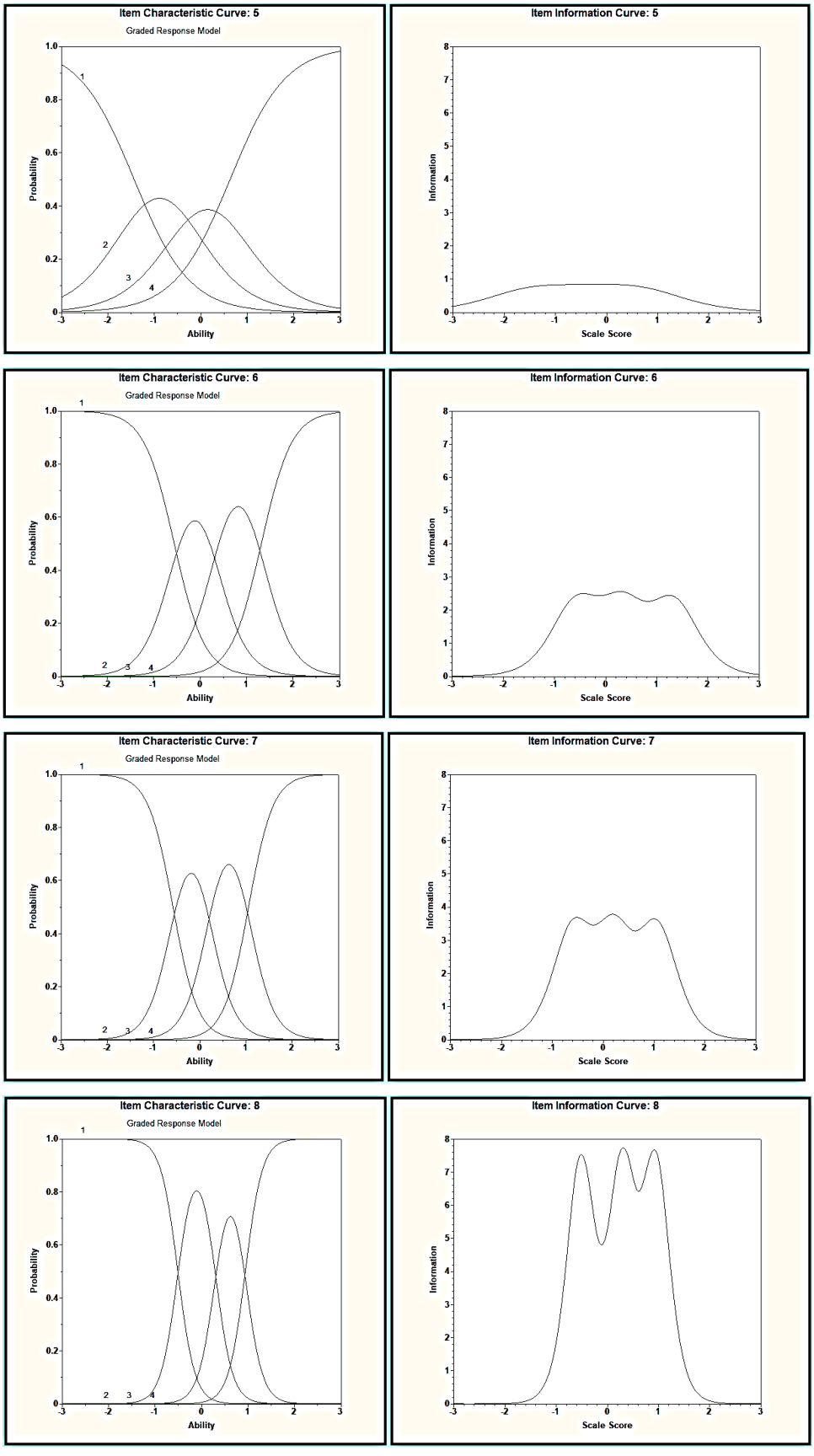


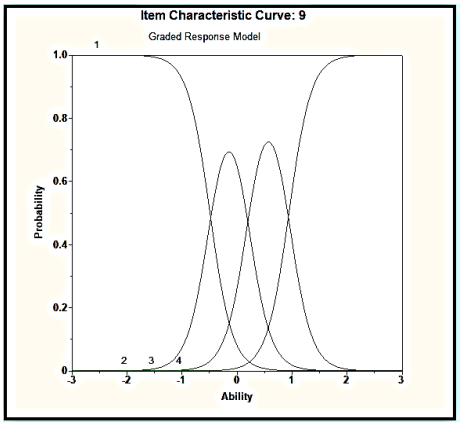


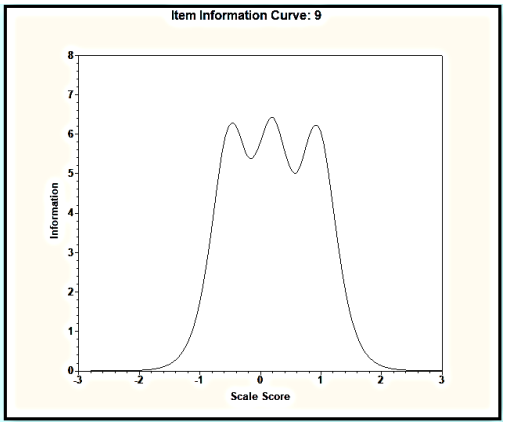


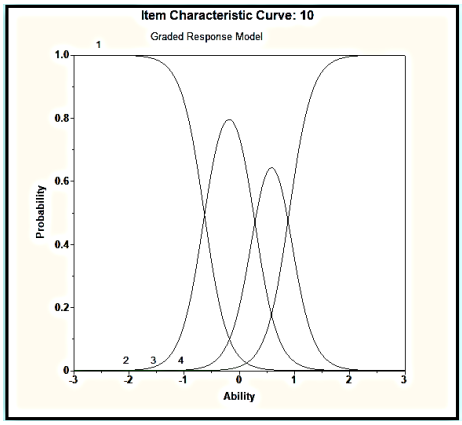


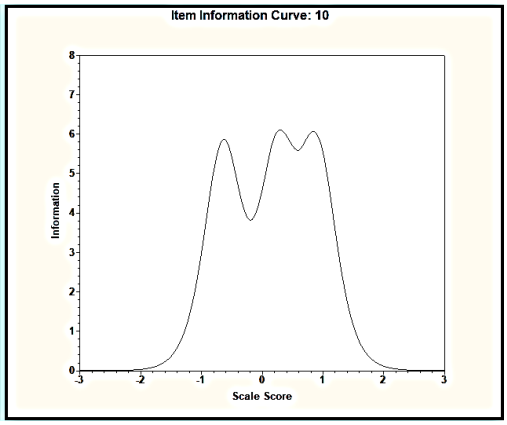


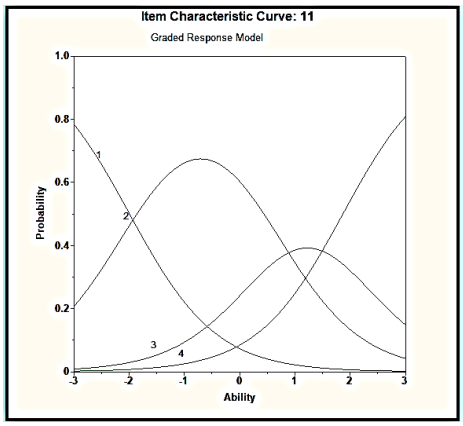

Item Characteristic Curve: 5
Graded Response Model
Probability
Ability
Item Information Curve: 5
Information
Scale Score
Item Characteristic Curve: 6
Graded Response Model
Probability
Ability
Item Information Curve: 6
Information
Scale Score
Item Characteristic Curve: 7
Graded Response Model
Probability
Ability
Item Information Curve: 7
Information
Scale Score
Item Characteristic Curve: 8
Graded Response Model
Probability
Ability
Item Information Curve: 8
Information
Scale Score

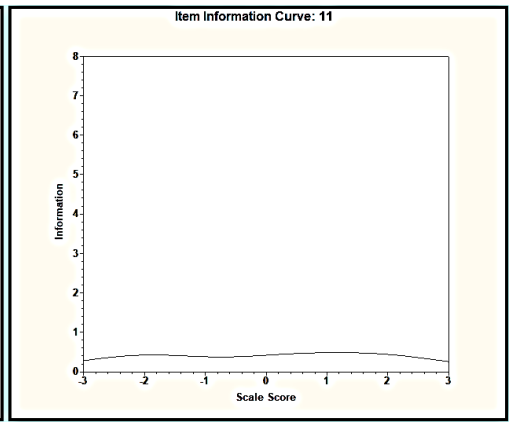

Item Characteristic Curve: 9
Graded Response Model
Probability
Ability
1
2
3
4


Item Information Curve: 9
Information
Scale Score

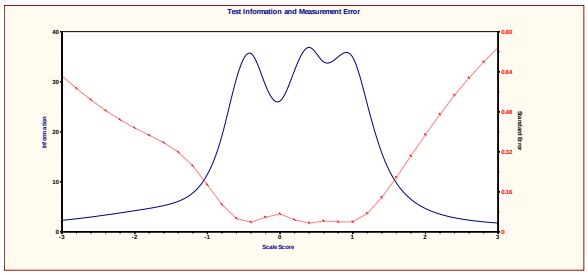

Item Characteristic Curve: 10
Graded Response Model
Probability
Ability
1
2
3
4


Item Information Curve: 10
Information
Scale Score

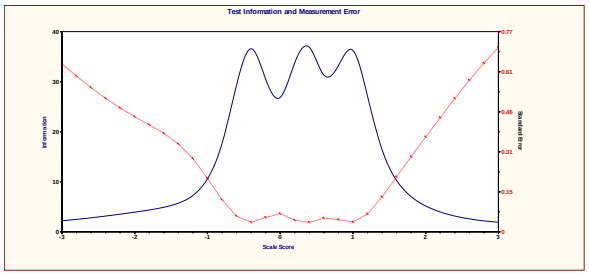

Item Characteristic Curve: 11
Graded Response Model
Probability
Ability
1
2
3
4


Item Information Curve: 11
Information
Scale Score

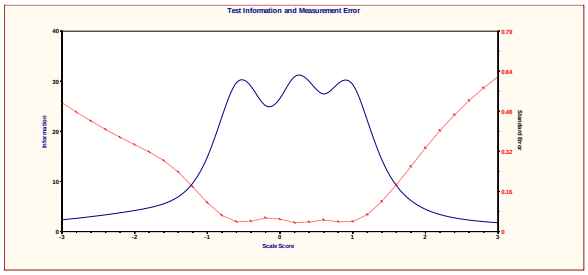
Appendix B. Test Information Functions of Five Raters



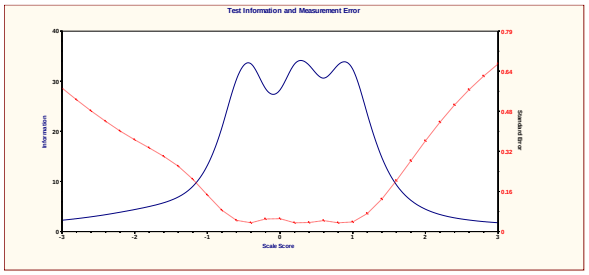
Test information function of the first rater



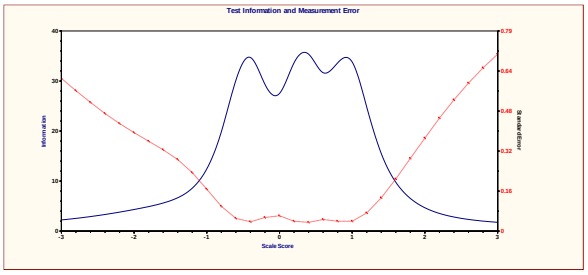
Test information function of the second rater



Test information function of the third rater



Test information function of the fourth rater



Test information function of the fifth rater