KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: May 13, 2016
Revision received: December 6, 2016
Accepted: January 23, 2017
OnlineFirst: February 28, 2017

www.estp.com.tr
DOI 10.12738/estp.2017.2.0321 ♦ April 2017 ♦ 17(2) ♦ 631–651

Research Article

A Comparison of Rubrics and Graded Category Rating Scales with Various Methods Regarding Raters' Reliability

C. Deha Doğan[1]
Ankara University

Müge Uluman[2]
Marmara University

Abstract

The aim of this study was to determine the extent at which graded-category rating scales and rubrics contribute to inter-rater reliability. The research was designed as a correlational study. Study group consisted of 82 students attending sixth grade and three writing course teachers in a private elementary school. A performance task was administered to students and student works were divided into two randomly. The teachers first conducted independent scoring on the works in group one by using the graded category rating scale, and then they scored the works in groups two with rubrics. Raters reliability was estimated by intra-class correlation coefficient, generalizability theory (G-theory) and many-facet Rasch model. The results indicated higher inter-rater reliability when graded category rating scale was used. Moreover qualitative data revealed that raters prefer using graded category rating scale in situations where they need to do quick scoring in short intervals. It is recommended that teachers use graded-category rating scale as a practical tool for quick scoring with the aim of determining student achievement with grades rather than giving detailed feedback.

Keywords

Inter-rater reliability • Generalizability theory • Many-Facet-Rasch Model • Analytical Rubrics • Graded Category Rating Scales

[1] **Correspondence to:** C. Deha Doğan (PhD), Department of Measurement and Evaluation, Faculty of Educational Sciences, Ankara University, Cebeci, Ankara 06590 Turkey. Email:dehadogan@gmail.com

[2] Department of Measurement and Evaluation, Faculty of Atatürk Education, Marmara University, Kadıköy, Istanbul 34722 Turkey. Email:mugeulumann@gmail.com

Citation: Doğan, C. D., & Uluman, M. (2017). A comparison of rubrics and graded category rating scales with various methods regarding raters' reliability. *Educational Sciences: Theory & Practice*, *17*, 631–651. http://dx.doi.org/10.12738/estp.2017.2.0321

The measurement of psychological attributes can be performed in two ways. One is to give examinees a number of stimuli (questions) assumed to stimulate the psychological attribute in focus; in this case, examinees respond to predetermined standard response categories depending on the magnitude of the attribute in themselves. This method, called direct self-report, simply covers tests in which examinees choose the correct answer from options available to them. In indirect self-report, examinees are similarly provided with a stimulus, but they are not offered standard response categories – they can create responses themselves (Erkuş, 2012). Examples of indirect self-report include open-ended questions and performance tasks which aim at measuring skills such as writing, critical thinking and creativity, where responses are structured by examinees themselves.

In tests which require examinees to select the correct answer, scoring can be carried out by scoring a correct answer as 1 and an incorrect answer as 0. In other words, there is no degree of correct answers. This is a condition that increases objectivity and reliability in scoring. However, as regards tests in which responses are constructed by examinees themselves, answers close to the correct answer are likely, as well as the correct answer. This situation adversely affects objectivity and reliability in scoring. In order to overcome the adverse effect, graded-category rating scales and rubrics are widely used for the scoring of such tools (Haladyna, 1997; Jonsson & Svingby, 2007; Rezaei & Lovorn, 2010).

Graded-Category Rating Scales

Graded-Category Rating Scales (GCRSs) are scoring tools containing the criteria regarding properties intended to be measured in student work along with success levels regarding these criteria. GCRSs offer dimensions for scoring and the range of scoring regarding student work (Haladyna, 1997). In this way, differentiation of the scores assigned by the raters can be prevented to a certain degree. However, since performance levels of success pertaining to the criterion are not defined, success measured as three points by one rater may be measured as two points by another. This situation can be considered as a major weakness of GCRSs. Conversely, their biggest advantage is their ease of preparation and quick scoring (Haladyna, 1997).

Rubrics

Rubrics are scoring tools used to define the criteria against which student work is evaluated, and the level to which their performance corresponds (Goodrich, 1996). The most distinctive aspect of rubrics in comparison to GCRSs is that they provide performance definition of each criterion for different levels of success. In other words, rubrics contain separate definitions made for each performance level regarding the criteria. In this context, rubrics have the advantage of giving feedback for students while enhancing objectivity in scoring for teachers. This ultimately contributes to a more standardized and objective determination, not varying from rater to rater.

Rubrics can be prepared for a certain scope (task-specific rubric). They can also be developed in order to score general skills such as writing and communication (generic rubric). In addition to these, there are two types of rubrics based on the structure of the tool (Haladyna, 1997). The first of these is the holistic rubric. In holistic rubrics, one single point is given to the student's entire performance and descriptions are available for all performance levels. Such rubrics are used in situations where small mistakes in student performance can be ignored and focus is placed on the whole performance (Arter & McTighe, 2001; Kutlu, Doğan, & Karakaya, 2010). The other type is the analytical rubric. Used more widely, the analytical rubric is a scoring tool that provides information about the achievement levels of student performance in various dimensions. Thus, it can provide a profile of the strengths and weaknesses of students in a certain area (Gronlund, 1998). While both types of rubric focus on performances in various dimensions and provide detailed definitions regarding performances at each level, analytical rubrics are a more reliable and functional tool than holistic rubrics. In this study, analytical rubrics were developed and used to collect the data.

A lack of performance definitions regarding the criteria can be seen as a major drawback of GCRSs overall. In contrast, rubrics have performance definitions but they are more time-consuming to prepare and apply. Still, it is an important question as to what extent the performance definitions of rubrics affect the reliability of scoring; and it is worth investigating to what extent scorings with GCRSs are reliable, as they do not include performance definitions. If they are at least as reliable as rubrics, teachers could be recommended to choose GCRSs to score student performance as they are more practical to prepare and use (It should be noted that rubrics are not used only to get reliable assessment results. On the other hand rubrics seem to have the potential of promoting learning and/or improve instruction. The main reason for this potential lies in the fact that rubrics make expectations and criteria explicit, which also facilitates feedback and self-assessment (Jonsson & Svingby, 2007). So rubrics are also used to promote learning and instruction processes. But this study focuses on the effects of rubrics on reliability and ignores the other advantages of rubrics). If rubrics provide more reliable scoring, it could be suggested that teachers use such tools for scoring even though they take longer to prepare. So, inter-rater reliability level must be determined and compared by various methods for both instruments.

Inter-rater reliability level can be determined by methods based on the classical test theory, generalizability (G) theory and item response theory. For determining agreement among raters, the Pearson product-moment correlation coefficient can be calculated, which determines the linear relationship between two raters based on the classical test theory. Though easy and practical to apply, this coefficient is affected by sample size, cannot be used in cases with more than two raters, and ignores similarities and differences in the scores assigned by the raters (it highlights the covariance between two variables independent on average points) (Goodwin, 2001; Güler & Gelbal, 2010a). In cases where the distribution does not meet the normality assumption, the Spearman–Brown correlation coefficient is used. Another method based on the classical test theory is Cohen's kappa coefficient, which aims to determine the level of agreement between two raters for data at categorical level. Overall, it is regarded as a more powerful statistical technique in comparison with simple percentage of agreement (Gravetter & Wallnau, 2007). One method based on the classical test theory is the weighted kappa coefficient, as a developed version of Cohen's kappa coefficient. While Cohen's kappa coefficient addresses raters' agreement on an individual examinee on a particular item with a holistic approach (scoring identical or not), the weighted kappa coefficient allows raters to give different weights on categories (Gisev, Bell, & Chen, 2013). Fleiss's Kappa coefficient, Kendall's W coefficient and the intra-class correlation coefficient are also used as statistics, again based on the classical test theory, which allow measurement of agreement level among three or more raters. In this study, to determine the inter-rater agreement on the basis of the classical test theory, an intra-class correlation coefficient (ICC) was used, since it has a more flexible structure for data analysis and the study includes three raters. ICC is an analysis technique which includes more than two raters used to assess different scoring sets, and in cases where there is missing data, it has three basic models: the one-way random effects model, the two-way random effects model, and the two-way mixed model. In the one-way random effects model, each item is scored by different sets of raters randomly selected from the population. In the two-way random effects model, each item is scored by all raters randomly selected from the population. Lastly, in the two-way mixed model, each item is scored by all raters in the population concerned. In essence, ICC is used for data at equal intervals and ratio scale level; it is also used for data at ranking level (Gisev et al., 2013).

Though raters are an important source of error, when open-ended items are used, there are also other sources of error which might interfere with measuring results. However, only one source of error is considered in reliability assessment methods based on the classical test theory. This does not allow the assessment of reliability based on different variance sources at the same time. Generalizability (G) theory allows calculation of a single reliability coefficient by assessing the errors that might come from a variety of sources of variability, such as raters, time, different forms of the test, tasks or items (Güler & Gelbal, 2010a). As an extension of the classical test theory, G theory can be specified as a model that includes multiple sources of error by benefiting from the analysis of variance (Brenann, 2011). In the present study, G-theory was also used for determining inter-rater agreement. The design used in this study, sources of variability and variance components are presented in "Data Analysis" below.

In essence, the classical test theory and the G theory are based on observed score. The same does not apply to the many-facet Rasch model (MFRM). In this model, as with other item response theory models, the latent trait of examinees is used as a probability of examinees' responses (Macmillan, 2000). MFRM is an extension of the basic Rasch model. The model estimates sources of variability of measurement by adding some parameters affecting examinee performance such as "raters" severity, task difficulty and any other sources of variability' (Iramaneerat, Yudkowsky, Myford, & Downing, 2008). By adding rater parameters into the measurement process, it becomes a useful instrument for estimating not only examinees' skills levels and the items' difficulty levels, but also the raters' severity levels as well (Linacre, Wright, & Lunz, 1990). The MFRM was also used in our study; detailed information regarding it is presented in 'Data Analysis' below.

During the literature review related to agreement between raters, one study was found comparing GCRS sand rubrics in terms of rater reliability. In the study carried out by Doğan and Yosmaoğlu (2015) in a private university's Health Sciences faculty, students' performance in a physiotherapy practical exam was scored independently by two raters using scoring rubrics and GCRSs. The rater reliability coefficients obtained from two tools were compared based on the results acquired from the different methods depending on classical test theory. It was discovered that there is higher agreement between the raters GCRS was used while scoring the items. In another study, Parlak and Doğan (2014) compared rater agreement in rating scales and rubrics by using methods based on the classical test theory, and concluded that the agreement is higher between raters in the case of rubrics. Another study was carried out by Büyükkıdık and Anıl (2015) which compares holistic and analytic scoring rubrics on the basis of G theory. They found that analytical rubrics have higher reliability.

In two other studies comparing the methods used for scoring, the effects of rubrics and answer keys on the same and different raters' reliability in grading essays were investigated. A comparison was made between essay scores given by different teachers at different times, with and without a rubric and answer key. As a result, it was found that there is significant difference between mean scores, and the time elapsing between ratings had an impact on the consistency of the scores. Comparison of essay scores rated by the same teacher at different times with and without a rubric and answer key also showed significant difference between mean scores, and the time between ratings played a distorting role on agreement (Kan, 2005a, 2005b).

In the literature, studies are available which investigate inter-rater reliability by using methods based on the classical test theory, G theory and item response theory. Besides, studies comparing methods used for determining inter-rater reliability based on different theories of measuring have an important place in recent research. These include studies comparing the methods based on the classical test theory, G theory, the many-facet Rasch measurement and the hierarchical rating model (Akın & Baştürk, 2010, 2012; Engelhard, 1994; Engelhard & Myford, 2003; Güler & Gelbal, 2010b; Güler & Teker, 2015; Iramaneerart, Myford, Yudkowsky, & Lowenstein, 2009; Iramaneerat et al., 2008; Linacre et al., 1990; Lynch & McNamara, 1998; Macmillan, 2000; Nakamura, 2000; Stenlund, 2013; Sudweeks, Reeve, & Bradshaw, 2004). Further details are not provided in relation to the above mentioned studies since the present study aims at comparing rubrics and graded-category rating scales used in scoring rather than comparing the methods used to determine inter-rater reliability.

The literature mainly consists of studies determining inter-rater reliability with different techniques or studies comparing different techniques. However, studies carried out to determine which scoring tools would increase raters' reliability are rare. Specifically, there is only one study which investigates the effects of GCRSs and rubrics on inter-rater reliability which is based on the classical test theory. Hence, this present study was carried out to determine and compare the potential role of GCRSs and rubrics on scoring reliability with methods based on the classical test theory, G theory and the MFRM. Within this framework, the study attempts to find answers to the following research questions.

1. In cases where rubrics and GCRSs are used, to what extent are raters reliable according to:
    a. the intra-class correlation coefficient;
    b. G theory; and
    c. the MFRM?
2. What are the raters' opinions regarding the feasibility and reliability of the scoring tools used in the study?

Methodology

In this section, the research model, study groups, data collection tools and data analysis methods used in this study are presented.

Research Model

This research is a correlational study. Correlational studies investigate the relationship between two or more variables without intervening in any way in these variables. Correlational research can be said to be an effective tool in uncovering the relationships between variables, determining the level of these relationships and providing essential tips for conducting higher-level research into them (Fraenkel & Wallen, 2006). In addition, quantitative findings were interpreted in relation to qualitative findings in the study.

During the study, participating students were assigned a performance task in writing, implemented under the teacher's control. Then, students' work was randomly divided in two. Firstly, the students' work in group 1 was scored by three teachers independently, using a graded scale. Then, those in group 2 were scored by the same teachers independently, using the rubric this time. After that, agreement was examined between scorings for both groups. The same teachers carried out both scorings with an eye to preventing differences arising from raters affecting the study result. The reason for scoring with the GCRS first (without performance definitions) was to prevent them from recalling the definitions while making their evaluation in the second group. Table 1 shows a model of the procedure implemented in this study.

Table 1
Model of the Study Procedure

	Rater	Scoring Tool	Scored Group
Stage 1	Rater 1 Rater 2 Rater 3	Rubric	Group 1
Stage 2	Rater 1 Rater 2 Rater 3	Graded-Category Rating Scale (GCRS)	Group 2

Study Group

Study data were collected in an elementary school within a private university. The data were taken from 82 students attending the sixth grade and three teachers teaching writing. Study participants were selected on a voluntary basis. The teachers participating in the study were selected if they already had five years' experience in their professional capacity. In order to reduce bias, raters and students who did not know each other had to participate in the study. To this end, the three teachers were selected from a different school to the students, and included two teachers from a state school and one from a private school. The implementation was carried out after briefing done with the raters.

Data Collection Tools

Study data were collected by using the scoring rubric and the GCRS developed by researchers for scoring the performance tasks in a writing class. There were five criteria in the scoring rubric and grading scale. Both tools included five performance levels. During the development process, two writing teachers with professional experience of five and six years respectively (who did not take part as raters in the study), two experts in assessment and evaluation, and one linguist, commented on the tools. They were then revised in the light of the feedback given. Before proceeding, the scoring tools were piloted so that researchers could take necessary actions. Five students took part in the pilot study. Essays written by these students were evaluated by two teachers who did not participate as raters in the study. Their feedback was also taken and both tools were finalized accordingly. The grading scale and scoring rubric developed in this study are given in Appendix 1 and Appendix 2.

Moreover to get a valid performance task (the writing task administered to participants) at first the writing skills aimed to assess was defined (learning outcomes). Second the content of the task (problem situation) was designed to assess those skills. Then the criteria in the scoring tools (rubric and GCRS) were matched with those skills (Kutlu et al., 2010). In the end there was harmony among the learning outcomes, content of the task and the criteria in scoring tools. So it was aimed to assess all the skills existed in learning outcomes without ignoring any of them. After the draft of the writing task was formed two experts examined the form and some revisions were made. According to the expert opinions one of the criteria (time to complete task) was removed because it assessed the skill that was not existed in the learning outcomes. On the other hand some of the instructions were revised so that students better understand what was expected to them. The process mentioned above was carefully followed for the validity of the performance task.

In addition, in order to determine the raters' views on the scoring process, a semi-structured interview form was developed by researchers. The form included questions concerning the raters' experience in feasibility, the reliability of the scoring tools, and their positive and negative aspects.

Data Analysis

To analyze the data obtained in this study, three different techniques were used. One of the techniques was the intra-class correlation coefficient (ICC), which is based on the classical test theory. ICC was preferred as it could determine the reliability of more than two raters, making the data analysis process flexible and able to be used with data at ordinal level. A two-way random effects ICC model was used in this study. The reason for selecting this model is that each item used in the study was scored by all raters selected randomly. For calculating the ICC, R software was utilized.

Generalizability (G) theory was also utilized in determining the raters' reliability. Student performances were scored by three raters. In this regard, items, raters and individuals (examinees) were taken as variability sources. As each examinee answered all of the items and each item was scored by all raters in the study, a fully crossed design was used. Besides individual, task and rater main effects, the interaction effects of individual-task, individual rater, task-rater and task-individual-rater were analysed in the study. Calculation of the main and interaction effects regarding the variability sources was conducted using the Edu G. 6.1 software.

Another technique for identifying raters' reliability is the many-facet Rasch model (MFRM). It is an extension of the basic Rasch model. It performs estimates by adding to the model any other sources of variability parameters affecting examinee performance besides rater severity and task difficulty (Iramaneerat et al., 2008; Linacre & Wright, 2004). The study included three sources of variability in MFRM, as in G theory. These were students, raters and item variability sources.

MFRM analysis provided information regarding the calibration map, standard error, separation index, separation reliability index and values of source of variability. Moreover, "fit statistics" are known in Rasch analysis as "infit" and "outfit" mean square values (Sudweeks et al., 2004). They provided information regarding the extent to which each examinee, rater and task matched with the values estimated by the model in the analysis. The FACETS program was utilized for the MFRM analysis. Qualitative data collected through interviews were analyzed with content analysis.

Findings

In this section, findings related to the research question are presented under separate headings according to the classical test theory, G theory and MFRM. To ensure internal consistency of the findings regarding G theory and MFRM, the parameters of examinees, raters and items were reported together. Nevertheless, the parameters regarding raters were highlighted, as they represent the main focus of the study.

Findings based on Classical Test Theory

This section shows an intra-class correlation coefficient (ICC) which indicates rater reliability in cases where scoring rubrics and GCRSs are used. Table 2 is below that gives the results of ICC for GCRS and Rubric.

Table 2
Results of ICC

Statistics	N	X	S	df_1	df_2	ICC	Sig.
GCRS							
1. Rater	205	2.82	1.28	204	408	.692	.00
2. Rater	205	2.87	0.96				
3. Rater	205	3.28	1.15				
Rubric							
1. Rater	205	2.52	1.33	204	408	.593	.00
2. Rater	205	2.94	1.01				
3. Rater	205	3.00	1.09				

Table 2 displays mean of the examinee's scores that assign from raters are relatively same for GCRS and rubric. Take in consideration means, first and second rater assign more similar scores with one another than the third one for GCRS. For rubric second and third rater assign more similar scores with one another than the first one. In the case of GCRSs, the ICC between raters was found as .692 (p < .01). For rubrics, the ICC was .593 (p < .01). These values prove higher inter-rater reliability as a result of using rubrics than GCRSs

Findings Based on G theory

In this section, the variability for examinees, raters, items and their interaction were based on G theory. In Table 3 below, G theory parameters of scores are given by three raters for 41 examinees on five criteria, using both the GCRS and the rubric. A fully crossed design was used which includes two sources of variability.

Table 3
Variances Estimated with G Theory and Total Variance Explanation Rates

	Variance source	df	Sum of squares	Mean squares	Variance	%
Graded-Category Rating Scale	b	40	358.37	8.96	0.51	37.7
	p	2	22.68	11.34	0.00	0.3
	m	4	45.23	11.31	0.01	0.9
	b x p	80	97.72	1.22	0.17	12.6
	b x m	160	83.71	0.52	0.05	3.6
	p x m	8	77.68	9.71	0.23	16.9
	b x p x m	320	120.59	0.38	0.38	28.0
Rubric	b	40	269.90	6.75	0.30	20.5
	p	2	29.34	14.67	0.02	1.3
	m	4	10.75	2.69	-0.06	0.0
	b x p	80	136.93	1.71	0.25	17.2
	b x m	160	167.65	1.05	0.19	13.3
	p x m	8	76.09	9.51	0.22	15.2
	b x p x m	320	150.31	0.47	0.47	32.4

Examining the related values, it can be seen that the estimated variance component for individual (b) main effects has the highest proportion (0.51) in the total variance (37.7%) when using the GCRS. We may infer that differences among individuals can be determined in this measurement. Examination of individual (b) main effects for the rubric shows that the variance component (0.30) and its proportion in the total variance (20.5%) is lower than in the GCRS. This finding might be interpreted as a better determination of differences between individuals when using GCRS.

For the GCRS, raters' main effects (p) have a low variance component (0.00) and percentage (0.3%). It seems that severity levels of scorings performed by raters for all examinees did not vary. For the rubric, as regards raters' main effects, the variance component (0.02) and percentage (1.3%) have low values, implying that raters have similar severity and leniency levels of scorings. Taking into consideration the values obtained for both scoring tools, it can be said that there are not huge differences between raters' scorings for all examinees. However, it can be suggested that differences between raters are smaller in the case of using a GCRS.

As regards the main effects of tasks (m), the rubric has a small variance component (-0.06) and a rather small proportion in the explained variance (0.0%). This finding might imply that difficulty levels of items do not differ. These values are also small for the GCRS; still, they are found to be higher than rubrics. Therefore, it can be argued that items have more similar difficulty levels when a rubric is used.

From the perspective of rater and examinee interaction effects (b x p), the variance component (0.17) and variance percentage (12.6%) have a higher value in the case of the GCRS. A higher variance component (0.25) and variance percentage (17.2%) were obtained in the case of the rubric. Taking into consideration the interaction effects of rater and examinee, it can be argued that raters scored some examinees with varying levels of severity and leniency in both cases where GCRSs and rubrics were used. In other words, rater and examinee interaction is not the same across the various raters (Sudweeks et al., 2004). Nevertheless, higher levels of variance components and percentages as a result of using rubrics might imply that the inter-rater difference is relatively smaller in the case of GCRSs.

From the perspective of interaction effects of examinee and task (b x m), the GCRS seems to have a lower variance component (0.05) and percentage (3.6%). Besides the variance component and percentage are respectively 0.19 and 13.3% for rubric. On the basis of this finding, it can be said that the relative status of examinees does not vary from one task to another. Higher values for the rubric might indicate that the relative status of examinees shows a larger variance between tasks in comparison to the GCRS.

Considering the interaction effects of rater and item as an indicator of whether scoring by raters is stable between items (p x m), it was seen that both the GCRS (variance component 0.23, variance percentage 16.9%) and the rubric (variance component 0.22, variance percentage 15.2%) have high levels of variance component and percentage. It seems that raters were not stable in scoring different items in relation to both scoring tools. Still, it could be suggested that the rubric allowed raters to perform relatively more stable scoring between items in comparison with the GCRS, as it has lower variance component and percentage.

From the examinee, rater and item interaction effect point of view (b x p x m), the GCRS is seen to have the second largest variance component (0.38) and percentage (28%). In the rubric, the interaction effect had the largest variance component (0.47) and percentage (32.4%). In the light of these findings, the level of random errors seems high for measuring with both tools. However, the interaction effect of examinee, rater and item is lower, which might imply that relatively fewer random measurement errors are involved in the GCRS than in the rubric.

In the study, task difficulty and student success were defined as assessments based on an absolute criterion rather than a relative criterion. Therefore, an absolute G coefficient (generally called as Phi coefficient) was considered. In cases where the GCRS was used, the Phi coefficient was calculated as 0.82, whereas it was found to be 0.63 in the case of using the rubric.

Obtained values demonstrate that the GCRS led to higher examinee main effects and G coefficients than the rubric, while rater main effects and examinee-rater interaction effects were lower. It seems from the findings that when a GCRS is used, inter-rater reliability is higher than when a rubric is used.

Findings based on Many-Facet Rasch Measurement

This section presents examinee, rater and item parameters for GCRS and rubrics' obtained by using the many-facet Rasch model (MFRM).

Parameters related to examinees. In order to decide whether or not the data fit to the model, standardized residuals were examined. For the GCRS, out of 615 data, there were three standardized residuals greater than -/+ 3 (0.65%) and eight standardized residuals greater than -/+ 2 (1.30%). On the other hand, the number of standardized residuals above -/+ 3 and -/+2 was five (0.81%) and 12 (1.95%) respectively for the rubric. Looking at these values, it can be suggested that research data fit to the MFRM.

For the variance source of examinee, in-fit and out-fit values, which provide model precision and the suitability of the performance displayed, were examined within a quality control fit criteria of 0.6 – 1.4 (Linacre et al., 1990; Nakamura, 2000). It was seen that in-fit and out-fit values for both GCRSs and rubrics were within the range, except for six examinees.

In the scope of examinee parameters, the separation indices were examined on the logit scale in order to identify the extent to which examinees vary from each other according to sources of variability. This value was found to be 3.46 for the GCRS, and examinees were placed in five skill levels as a result of the formula (4G+1)/3 (Lee & Kantor, 2003). The seperation index for the rubric was found to be 2.49, where examinees were placed in four skill levels. Interpreted in a similar way to the KR-20 and Cronbach's Alpha value, the separation reliability index was found to be 0.92 and 0.86 for the GCRS and the rubric respectively. This value predicts the reliability of the test in terms of internal consistency (Bond & Fox, 2001). Though both tools are reliable according to this value, the GCRS seems to have higher reliability than the rubric. In other words, students' writing skills can be scored with higher reliability using a GCRS than a rubric. In order to find out if the variability is significant, the seperation reliability index and the null hypothesis was tested by using chi-square (GCRS: $X^2 = 454.6$, df = 40, $p = .00$; rubric: $X^2 = 240.5$, df = 40, $p = .00$) and rejected as a result. According to this finding, the variability seems statistically significant for both of the tools used.

Parameters related to items. An analysis was carried out on in-fit and out-fit values of the criteria in GCRSs and rubrics. Except for the first criterion, in-fit and out-fit values were found to be within the quality control fit criteria for both instruments (GCRS: 1.7, rubric: 1.8). This criterion was seen to have misfit with the other criteria due to a variance of more than 70% for the GCRS. Similarly, a misfit was found for the rubric with a variance of over 80%.

Parameters regarding the criteria (items) that were used to assess students' writing skills were analyzed. RMSE values yielding the standard error were reported as 0.12 and 0.10 for the GCRS and the rubric respectively. Since these values are quite close to each other, it can be stated that standard error of criteria was quite low for both instruments. For the GCRS, the separation index for items was 3.78, which is above the recommended initial value (2.00) (Nakamura, 2002). This result indicates that the criteria could distinguish students with different levels of skill. The separation reliability index coefficient was found to be 0.93, which suggests that the criteria in the study seem highly reliable for determining students' writing skills. For the rubric, the separation index was found to be 1.23, which is below the initial value. Compared to the GCRS, the criteria included in the rubric seemed to be less competent at determining students with different levels of skill. Its separation reliability index coefficient was 0.60, which demonstrates that rubrics are less reliable than GCRSs at determining students' writing skills. For both tools, the null hypothesis was tested with chi-square (GCRS: $X^2 = 74.6$, df = 4, $p = .00$; rubric: $X^2 = 12.5$, df = 4, $p = .01$) and rejected consequently. Variance between the criteria could be said to be significant for both instruments. Item Parameters for GCRS and rubric are given in Appendix 3

Parameters related to raters. Scores given by raters for students' responses are given in Table 4 below for detailed inquiry, as required by the aim of our study.

Table 4
Results of Raters' Severity and Leniency

				Rater severity		*In-fit*		*Out-fit*	
	Rater no.	Rater av.	Rater total r	Logit measure	S.E.	Squares av.	Z std.	Squares av.	Z std.
Graded-Category Rating Scale	1	2.9	2.95	.22	.09	1.0	0.0	1.0	0.00
	2	2.9	2.95	.22	.09	1.0	0.0	1.0	0.00
	3	3.3	3.39	-.45	.09	1.0	0.0	0.9	0.00
	Mean	3.0	3.09	.00	.09	1.0	0.0	1.0	0.0
	SS	0.2	0.21	.32	.00	0.0	0.2	0.0	0.4
	RMSE (Model) = 0.09, SS = 0.30, Seperation Index = 3.40, Reliability = 0.92 Fixed (all same) chi-square = 36.9, Sd = 2, *p* = .00								
				Rater severity		*In-fit*		*Out-fit*	
	Rater no.	Rater av.	Rater total r	Logit measure	S.H.	Squares av.	Z std.	Squares av.	Z std.
Rubric	1	2.5	2.46	.35	.08	1.1	1	1.1	0
	2	2.9	2.96	-.14	.07	0.8	-1	0.9	-1
	3	3.0	3.03	-.21	.08	1.0	0	1.0	0
	Mean	2.8	2.81	.00	.08	1.0	0.0	1.0	-0.1
	SS	0.2	0.25	.25	.00	0.1	1.4	0.1	0.9
	RMSE (Model) = 0.08, SS = 0.24, Separation Index = 3.17, Reliability = 0.91 Fixed (all same) chi-square = 33.0, Sd = 2, *p* = .00								

Table 4 displays raters in decreasing order of severity towards leniency. Despite differing logit values of raters for both the GCRS and the rubric, they are in the same order. Rater 1 was listed as the most sever rater (GCRS: 0.22; rubric: 0.35), while rater 3 became the most lenient one (GCRS: -0.45; rubric: -0.21). Except for the extreme outlier values, the RMSE value showing the all data standard error was found to be 0.09 for GCRSs and 0.08 for rubrics. Close values obtained from both tools suggest that standard error level is low. In other words, the two instruments are found to be quite similar as regards standard errors relating raters' severity towards leniency.

When raters' in-fit and out-fit values were examined, they were seen within the quality control fit criteria for both the GCRS and the rubric (0.6-1.4). Hence, it could be said that while using both GCRSs and rubrics, raters scored consistently among themselves and with each other.

As an indicator of undesired inter-rater variance, the separation index (Sudweeks et al., 2004) was found to be 3.40 for the GCRS, with a separation reliability index level of 0.92. Though a specific upper limit is not available for the separation index, values close to 0.00 are preferable (Myford & Wolfe, 2004). Therefore, there might have been differences between raters during scoring or there might have been error in scores arising from raters. The separation index and reliability null hypothesis was tested with chi-square ($X^2 = 36.9$, df = 2, p = .00) and rejected. This finding might refer to statistical differences between raters' severity and leniency level. For the rubric, the separation index and separation reliability index was calculated as 3.17 and 0.91 respectively. The hypothesis that there is no significant difference between the separation reliability index and constant effect level of rater severity and leniency was tested with chi-square ($X^2 = 33.0$, df = 2, $p = .00$) and rejected in the end (Nakamura, 2002). Hence, it could be suggested that there were differences between raters during scoring in the rubric as in the GCRS. Also, scores given to items are dependent not only on the quality of the item, but also the raters. On the other hand, raters' standard Z points were similar in the rubric and these points were identical for the GCRS. For both of the tools, there was found to be a difference slightly above one unit of logit from a severity and leniency perspective. In this regard, the differences between raters seemed at a tolerable level, implying that raters were similar in terms of scoring severity and leniency (Lee & Kantor, 2003). Taking into account the parameters obtained in this study, rater reliability in GCRSs was found to be relatively higher than in rubrics.

Findings from Interviews

After scoring was completed, interviews were held with raters in order to obtain their opinions regarding the scoring using the two tools. As a result of analysis of face-to-face and telephone interviews, three main themes were identified. These themes are described below.

1. *Practicability in scoring.* Raters pointed out that they completed the scoring quickly due to the end of term and their workload. They reported that the GCRS was a more practical tool. Below are some examples from raters' statements:

   Rater 1: It was a busy period while I was doing the scoring. So, I had to do it fast. In that process, the GCRS was a more useful tool than the rubric.

   Rater 3: As the rubric includes definitions for every criterion, scoring took longer. I could do easier and quicker scoring with the GCRS.

2. *Fatigue during scoring.* Raters reported more tiredness while using the rubric and emphasized that their scoring behavior was affected by that. Below are some examples from raters' statements:

   Rater 2: To tell the truth, it was so tiring for me to refer to the definitions related to the criteria every time while I was scoring using the rubric. In some cases, I did the scoring without reading the definitions for the criteria.

   Rater 3: With the rubric, it was tiring to read definitions for each criterion. If I had done the scoring when I was less busy, I could have done a better measurement with the rubric.

3. *Feedback.* Overall, raters proposed that the rubric could provide more effective feedback for students while emphasizing that more time is needed for effective use of the tool. Below are some examples from raters' statements:

   Rater 1: The GCRS allows quicker scoring. Of course, students can be given more effective feedback with the rubric. But to achieve this and use the rubric more effectively, more time should be allocated for scoring each student's work. This becomes unlikely because of the workload that must be finished and the high number of students.

   Rater 2: Because there are definitions in the rubric, students can see their weaknesses better. Also, I can see the level of each student more clearly. But this requires more effort ... If my goal is to grade my students, I think the GCRS is more practical.

According to the analysis of the qualitative data, the scoring tool preferred by raters could be explained by the purpose of evaluation and circumstances under which evaluation took place. It could be suggested that their behaviors during scoring were guided by these factors.

Conclusion and Discussion

In this study, a writing task was assigned to sixth-graders, and their essays were scored by three raters using the GCRS and the rubric. Inter-rater reliability coefficients were identified by means of intra-class correlation coefficients based on the classical test theory, G theory and MFRM. Results obtained from the three different methods indicated higher inter-rater reliability in cases where the GCRS was used. The finding seems to contradict a similar study comparing GCRSs and rubrics (Doğan & Yosmaoğlu, 2015).

Two of the themes in the interviews were identified as 'quick scoring' and 'fatigue during scoring'. Raters pointed out that they could complete the scoring in a quicker and more practical way with the GCRS, while scoring with the rubric was more tiring. Another concept emphasized in the interviews was 'feedback'. The raters said that they did not know the students and did not need to give feedback. This might have caused the raters to take the scoring less seriously while using the rubric. Likewise, they said that GCRSs were more practical and useful if they were not required to provide feedback.

It seems that in order to decide the assessment tool (GCRS or rubric) to be used, the purpose of the assessment (summative or formative) and practical conditions (workload, available time to score, etc.) are decisive. As an example, rubrics are not recommended for summative assessment due to exhausting scoring and a lower level of inter-rater reliability. Instead, GCRSs should be used because of the advantages of higher rating reliability and practicability in scoring.

Conversely, rubrics seem more logical for formative assessment because they allow teachers to give more effective feedback to students, despite having a lower level of inter-rater reliability. The reason is that the main purpose of these types of assessment is to contribute to student development by means of feedback alongside reliable scoring. The raters in our study pointed out that they would prefer using a rubric if they were required to give feedback.

In the light of the study findings, we make the following recommendations.

For teachers:

- a GCRS would be useful in cases where summative assessment and quick and reliable scoring are required; and
- a rubric could be used in cases where formative assessment and giving effective feedback to students are required.

For researchers, it could be useful to:

- conduct similar studies with more raters;
- employ nested patterns;
- use the hierarchical scoring model in determining inter-rater reliability; and
- compare analytical and holistic scoring rubrics in terms of inter-rater reliability.

References

Akın, Ö., & Baştürk, R. (2010). Assessment of research assignment by many-facet Rasch measurement approach. *Journal of Measurement and Evaluation in Education and Psychology, 1*(1), 51–57.

Akın, Ö., & Baştürk, R. (2012). The evaluation of the basic skills in violin training by many-facet Rasch model. *Pamukkale University Journal of Education, 31*, 175–187.

Arter, J. A., & Mctighe, J. (2001). *Scoring rubrics in the classroom*: *Using performance criteria for assessing and improving student performanc*e. Thousand Oaks, CA: Corvin Press.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Brennan, R. L. (2001). *Generalizability theory.* Iowa City, IA: ACT Publications.

Büyükkıdık, S., & Anıl, D. (2015). Investigation of reliability in generalizability theory with different designs on performance based assessment. *Education and Science, 40*(177), 285–296.

Doğan, C. D., & Yosmaoğlu, B. (2015). The effect of the analytical rubrics on the objectivity in physiotherapy practical examination. *Türkiye Klinikleri Journal of Sports Science, 7*(1), 9–15.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112.

Engelhard, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and composition program with a many-faceted Rasch model. *ETS Research Report Series, 1*, i–60. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01893.x

Erkuş, A. (2012). Ölçmecilere bulaşan yeni yanılgılar ve yanlışlar [Recent errors and mistakes of measurement scholars]. *Elementary Education Online, 11*(1) 1–9.

Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education*. New York, NY: McGraw-Hill International.

Gisev, N., Bell, J. S., & Chen, T. E. (2013). Interrater agreement and interrater reliability: Key concepts, approaches and applications. *Research in Social and Administrative Pharmacy, 9*, 330–338.

Goodrich, H. (1996). *Students' self-assessment: At the intersection of metacognition and authentic assessment* (Doctoral dissertation, Cambridge, MA: Harvard University).

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, *5*(1), 13–14.

Gravetter, F. J., & Wallnau, L. B. (2007). *Statistics for the behavioural sciences* (7th ed.). Belmont, Canada: Thomson Wadsworth.

Gronlund, N. E. (1998). *Assessment of student achievement*. Boston, MA: Allyn & Bacon.

Güler, N., & Gelbal, S. (2010a). Studying reliability of open-ended mathematics items according to the classical test theory and generalizability. *Educational Sciences: Theory & Practice, 10*, 989–1019.

Güler, N., & Gelbal, S. (2010b). A study based on the classical test theory and many facet Rasch model. *Eurasian Journal of Educational Research, 38*, 108–125.

Güler, N., & Teker, G. T. (2015).The evaluation of rater reliability of open-ended items obtained from different approaches. *Journal of Measurement and Evaluation in Education and Psychology, 6*(1), 12–24.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston, MA: Allyn & Bacon.

Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education*, *13*(4), 479–493.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequence. *Educational Research Review, 2*, 130–134.

Kan, A. (2005a). The effect of using grading scale and response key to (same) grader's reliability. *Eurasian Journal of Educational Research, 19,* 166–167.

Kan, A. (2005b). The effect of using grading scale and response key to (different) grader's reliability. *Eurasian Journal of Educational Research, 19,* 207–219.

Kutlu, Ö., Doğan, D., & Karakaya, İ. (2010). *Öğrenci başarısının belirlenmesi performansa ve portfolyaya dayalı durum belirleme* [Assessing student achievement: Performance based and portfolio based assessment]. Ankara, Turkey: Pegem Akademi Press.

Lee, Y.-W., & Kantor, R. (2003). *Investigating differential rater functioning for academic writing samples: An MFRM approach.* Educational Testing Service.

Linacre, J. M., Wright, B. D., & Lunz, M. E. (1990). *A facets model for judgmental scoring*. MESA Memo, 61. Chicago, IL: MESA.

Linacre, J. M., & Wrigth, B. G. (2004). Construction of measures from many-facet data. In E. M. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory models and applications* (pp. 296–321). Maple Grove, MN: JAM Press.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180.

Macmillan, P. D. (2000). Classical, generalizability, and multi-faceted Rasch detection of inter-rater variability in large, sparse datasets. *The Journal of Experimental Education*, *68*(2), 167–190.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*, 189–227.

Nakamura, Y. (2000). Many-facet Rasch based analysis of communicative language testing results. *Journal of Communication Students*, *12*, 3–13.

Nakamura, Y. (2002). Teacher assessment and peer assessment in practice. *Educational Studies, 44*, 203–215.

Parlak, B., & Doğan, N. (2014). Comparison of answer key and scoring rubric for the evaluation of student performances. *Hacettepe University Journal of Education, 29*(2), 189–197.

Reza Rezaei, A., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*, 18–39.

Stenlund, T. (2013). Agreement in assessment of prior learning related to higher education: An examination of inter-rater and intra-rater reliability. *International Journal of Lifelong Education, 32*(4), 535–547.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*(3), 239–261.

Appendix 1

Graded Category Rating Scale Used to Assess The Students' Performance

	Very poor (1)	Poor (2)	Not bad (3)	Good (4)	Very Good (5)
Correct use of derived words					
Completing the missing part of the passage meaningfully					
Correct use of spelling and grammar					
Neatness of page layout					
Conformity of the title and passage completed.					

Appendix 2

Rubric Used to Assess the Students' Performance

	Very poor (1)	Poor (2)	Not bad (3)	Good (4)	Very Good (5)
Use of Derivative Words	2 or less than 2 derivate words used properly	3 or 4 derivate words used properly	5 or 6 derivate words used properly	7 to 9 derivate words used properly	10 derivate words used properly
Completing the missing part of the passage	Development and conclusion sections were not clearly inserted and there is no cohesion.	Development and conclusion sections were inserted, but there is no cohesion. Cohesion between paragraphs is weak.	Development and conclusion sections were inserted meaningfully, But Cohesion between paragraphs is partially good.	Development and conclusion sections were inserted meaningfully Cohesion between paragraphs is good but there is little mistakes.	Development and conclusion sections were inserted meaningfully. Cohesion between paragraphs is very good.
Spelling and grammar	There are more than 8 spelling and grammar mistakes in overall essay.	There are 7 or 8 spelling and grammar mistakes in overall essay.	There are 5 or 6 spelling and grammar mistakes in overall essay.	There are 3 or 4 spelling and grammar mistakes in overall essay.	There are maximum 1 or 2 spelling and grammar mistakes in overall essay.
Page layout	There are important deficiencies in page layout. Writing is not legible; indents and line breaks are not aligned.	There are important deficiencies in page layout. Writing is a little legible; indents and line breaks are not aligned.	Page layout is partially neat and clean. Writing is partially legible; indents and line breaks are partly aligned.	Page layout is neat and clean. Writing is legible; but indents and line breaks are not completely aligned.	Page layout is partially neat and clear. Writing is legible; indents and line breaks are aligned.
Title	There is no relation between title and content.	There is partial relation between title and content. The title is ordinary.	There is relation between title and content but the title is ordinary	There is relation between title and content; but title is partly original.	There is relation between title and content; and title is original

Appendix 3

Item Parameters for GCRS and Rubric

GCRS Item Parameters Table

			Item Difficulty		*In-fit*		*Out-fit*	
Item no.	Item av.	Item total r	Logit measure	S.E.	Squares av.	Z std.	Squares av.	Z std.
1	2.7	2.73	.51	.11	1.8	5.0	1.7	4.0
2	2.8	2.84	.37	.11	0.8	-1.0	0.8	-1.0
3	2.9	3.02	.13	.11	0.6	-3.0	0.7	-2.0
4	3.2	3.30	-.31	.12	0.7	-2.0	0.7	-2.0
5	3.4	3.54	-.71	.12	1.0	0.0	1.0	0.0
Mean	3.0	3.08	.00	.12	1.0	-0.4	1.0	-0.4
SS	0.3	0.30	.45	.00	0.4	3.0	0.4	2.7

RMSE (Model) = .12, SS = .44, Separation Index = 3.78, Reliability = .93
Fixed (all same) chi-square = 74.6, Sd = 4, *p* = .00

Rubric Item Parameters Table

			Item Difficulty		*In-fit*		*Out-fit*	
Item no.	Item av.	Item total r	Logit measure	S.E.	Squares av.	Z std.	Squares av.	Z std.
1	2.8	2.77	.04	.10	1.8	5.0	1.8	5.0
2	2.8	2.75	.06	.10	0.6	-4.0	0.6	-4.0
3	3.1	3.12	-.30	.10	1.0	0.0	1.0	0.0
4	2.7	2.72	.09	.10	0.7	-3.0	0.6	-3.0
5	2.7	2.71	.11	.10	1.0	0.0	1.0	0.0
Mean	2.8	2.81	.00	.10	1.0	-0.5	1.0	-0.5
SS	0.1	0.15	.15	.00	0.4	3.4	0.4	3.4

RMSE (Model) = .10, SS = .12, Separation Index = 1.23, Reliability = .60
Fixed (all same) chi-square = 12.5, Sd = 4, *p* = .01