iejee | INTERNATIONAL ELECTRONIC JOURNAL OF ELEMENTARY EDUCATION

March 2020, Volume **12**, Issue **4**, **349-358**

Investigation of Rater Tendencies and Reliability in Different Assessment Methods with Many Facet Rasch Model

Duygu Koçak*

Received: 12 September 2019
Revised: 18 March 2020
Accepted: 28 March 2020
ISSN: 1307-9298

www.iejee.com

DOI: **10.26822/iejee.2020459464**

Abstract

One of the most commonly used methods for measuring higher-order thinking skills such as problem-solving or written expression is open-ended items. Three main approaches are used to evaluate responses to open-ended items: general evaluation, rating scales, and rubrics. In order to measure and improve problem-solving skills of students, firstly, an error-free measurement process should be performed. Errors caused by raters such as bias, high or low tendency to score is a common problem in the evaluation of open-ended items as they adversely affect the accuracy of decisions to be made. This study utilized open-ended items to evaluate the raters' tendencies in terms of general evaluation, rating scale, and rubric conditions. The raters' behaviours in each assessment method and their opinions about the assessment methods were determined. The participants of the study consisted of 12 different mathematics teachers and the Many Facet Rasch Model was adopted for the analyses. The scoring reliability of each method was estimated. The findings of the rating scale revealed that the raters had a more homogeneous scoring tendency. In addition, while the majority of raters stated that they prefer to use a rubric, they also stated it is the most difficult method to use.

Keywords: Many Facet Rasch Model, Problem-Solving, Rater Reliability, Rater Tendency, Rating Scale, Rubric

Introduction

The quality of assessment, monitoring, and evaluation processes are directly related to the quality of the measurement tools used in these processes. The quality of these tools, which entails the ability to measure as far as possible and without errors, is determined by the quality of the items. Different item types have been developed to measure learning at different cognitive levels during the education process (Çıkrıkçı, 2010). The two basic item structures used are multiple-choice items that students choose to respond to and constructed response items that students construct themselves (Crocker & Algina, 1986; Roid & Haladyna, 1982). When deciding which item to use, item type which is more suitable for the feature to be measured, is recommended (Kastner & Stangla, 2011; Popham, 2008; Rodriquez, 2002; Roid & Haladyna, 1982). Therefore, the main factor to be considered is the cognitive level of the feature to be measured. Especially in classroom assessments, where multiple knowledge and skills are wanted to be measured at different cognitive levels, different item structures can be used together. Many large-scale national and international evaluation studies such as National Assessment of Educational Progress (NEAP), Scholastic Aptitude Test (SAT), Trends in International Mathematics and Science Study (TIMMS), Programme for International Student Assessment (PISA) include both multiple-choice and open-ended items (DeCarlo, Kim, & Johnson, 2011; Kim, 2009; Mariano, 2002).

In some cases, the use of open-ended items is a necessity. For example, using multiple-choice items to measure problem-solving skills only serves for "finding the right result" step of problem solving skills. However, problem-solving process is also crucial for problem solving skill. Therefore, open-ended items are often needed in mathematics classes. Open-ended items are used in situations where students are asked to form their own answer, such as problem solving (Kastner & Stangla, 2011; Messick, 1994; Park, 2017; Rodriquez, 2002; Roid & Haladyna, 1982). Open-ended items are beneficial if students need to plan and configure the answer to the question at hand (Haladyna, 1997). Using open-ended items, more in-depth measurements of students' knowledge and skills can be conducted (Pollack, Rock, & Jenkins, 1992; Rodriquez, 2002). In addition, incomplete and inaccurate learning can be detected (Cooper, 1981). With open-ended items, students' responses can be obtained in a similar way to the behaviours that students should exhibit in real life (Popham, 2008). Because of these advantages, open-ended items are frequently used in classroom measurement and evaluation activities.

In addition to the advantages of open-ended items, there are also negative aspects such as the difficulty of scoring (DeCarlo, 2005, 2010; DeCarlo et al., 2011; Linacre, 2003; Popham, 2008; Wang, 2012). While two different raters will give the same multiple-choice test item the same score, it may not always be possible for two different raters to evaluate the same open-ended test item at the same score (Haladyna, 1997), because open-ended items do not have a clear and a single correct response as in multiple-choice items. In the scoring process of these items, there is more than one rater who uses general evaluation, a rating scale, or a rubric. In general evaluation approach, evaluation is made according to the criteria determined by the rater. Therefore, it can result in errors such as severity, leniency, or bias caused by the rater. Rating scales, on the other hand, provide raters with basic assessment criteria while not completely preventing the leniency and severity of raters. Rubrics provide raters with both assessment criteria and explanations of those criteria. Therefore, it is an evaluation method that prevents errors caused by raters more than the others. Accordingly, the evaluation method determined will significantly reduce errors caused by raters, although not completely eliminating them.

Evaluations conducted with more than one rater will increase the accuracy in determining student achievement (Mariano, 2002). It is essential for validity and reliability that the scores given by raters during evaluation processes are as accurate and as fair as possible (Linacre, 1994). However, although it is tried to be prevented by the choice of evaluation method, rater errors such as rater generosity, inconsistency, and bias

*Correspondance Details: Duygu Koçak, Department of Educational Sciences, Faculty of Education, Alanya Alaattin Keykubat University, Turkey. E-mail: duygu.kocak@alanya.edu.tr

occur (Myford & Wolfe, 2003; Donoghue & Hombo, 2000). Especially when there are differences between the scores given by different raters, a situation that is often difficult to solve, it is necessary to determine how raters differ (Linacre, 1990). The fact that raters give different scores indicates that rater reliability and objectivity are low. Rater reliability is defined as the degree of consistency between the scores of two or more raters regarding different individuals and different items (Aiken, 2000; Anastasi & Urbina, 1997). It is crucial to consider the presence of rater effects, especially when using open-ended substances (Kim, 2009; Linacre, 1994). Under different theories for determining rater effects, there are techniques such as generalizability theory, Cohens 'Kappa coefficient, and Fleiss Kappa coefficient, which are used in the literature. Another technique used to determine rater effect is the Many Facet Rasch Model.

The Many Facet Rasch Model incorporates the rater parameter, allowing raters to estimate the severity level simultaneously (Linacre, Wright, & Lunz, 1990). In this way, bias caused by raters in measurements of students and items are eliminated (Linacre, 1989; Sudweeks, Reeve, & Bradshaw, 2004). In this model, four factors that are generally thought to affect student scores are defined. These factors are student level, item or task difficulty, rater severity, and assessment tool (Linacre et al., 1990; Linacre & Wright, 2004). If necessary, other influencing factors can be added to the model. The most important advantage of the model is that it treats different raters as a source of variability. Rater severity or leniency means that any scores that are given by a rater are systematically higher or lower than the average scores given by other raters. This is also referred to as rater effect or rater error (Engelhard & Myford, 2003). This model includes rater severity levels. An effective rater is an individual who can always score with the same tendency and share a common understanding of the rating scale with other raters. In other words, no matter which rater scores a student, the score should always have the same relationship with all raters. This indicates objectivity (Linacre, 1994).

In the literature, there are many studies conducted using the Many Facet Rasch Model (Akın & Baştürk, 2012; Atılgan, 2005; Baştürk, 2010; Engelhard, 1994; Engelhard & Myford, 2003; Iramaneerart, Myford, Yudkowsky, & Lowenstein, 2009; Linacre et al., 1990; Nakamura, 2000; Nakamura, 2002) and directly examining the Many Facet Rasch Model (Casabianca & Junker, 2013, 2014; DeCarlo 2010; DeCarlo et al., 2011; Iramaneerat, Yudkowsky, Myford, & Downing, 2008; Kim 2009; Lynch & McNamara, 1998; Mariano, 2002; Patz, Junker, & Johnson, 2000; Patz, Junker, Johnson, & Mariano, 2002; Sudweeks et al., 2004; Verhelst & Verstralen, 2001; Wilson & Hoskens 2001). These studies are aimed at revealing the effect of the scoring category and the number of raters on the reliability of the measurements. Junker and Patz (1998), DeCarlo et al. (2011), Donough and Hombo (2000), Mariano (2002), Patz et al. (2002) stated that multi-category scoring would increase the accuracy of scoring. Junker and Patz (1998) stated that more accurate measures of student achievement could be obtained by using higher number of raters rather than giving students more items. Lunz and Schumacker (1997) found that task difficulty was useful in scoring. Alharby (2006) compared two different approaches in scoring (holistic and analytical rubric) and examined the reliability of the measurements. They found that the holistic approach was a better fit for analytic approach. Sebok (2010) stated that it is advantageous to use the Many Facet Rasch Model model with small samples and missing data, and it is the most appropriate method when individual evaluation is desired.

There are also studies that compare different assessment methods in the literature like Doğan and Uluman (2017), Boztunç-Öztürk, Şahin and İlhan (2019), Çetin and Kelecioğlu (2004), Ömür and Erkuş (2013), Akın and Baştürk (2012), Engelhard (1994), Sudweeks, Reeve and Bradshaw (2004), Özbaşı and Arcagok (2019). These studies investigated different scoring methods with G theory and/or Many Facets Rasch model. Doğan and Uluman (2017) determined the extent to which graded-category rating scales and rubrics contribute to inter-rater reliability. They estimated raters reliability by intraclass 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. Another example study conducted by Alharby (2006) aimed to determine the reliability of the measurements and rater tendencies to give scores obtained by three different assessment methods where different assessment methods were compared using two different rubric types. In the present study, the scoring tendencies of the raters were evaluated individually using the Many Facet Rasch Model.

In the Many Facet Rasch Model, values are generated for a measurement (logit estimation obtained from the analysis), a standard error (information on the precision of the logit estimation) and compliance indicators (information on how well the data fit into the model) for all elements of all variability sources (Engelhard & Myford, 2003). The validity of parameter estimations is obtained by statistical quantification of the fit of the model with the data (Wright & Masters, 1982). Therefore, the reliability and the validity of raters can be estimated in addition to rater trends in the model. Especially the prediction of teacher tendencies is the most significant advantage of the model since raters' severity/leniency significantly affects the reliability and the validity of the measurements. In this respect, the present study aimed to examine the change of rater tendencies according to general evaluation, rating scale, and rubric. For this purpose, the change of raters' tendencies and rater reliability to the assessment method are estimated with the Many Facet Rasch Model.

Methodology

Research Design

Qualitative and quantitative data are used together in the research. In this respect, the research is mixed model research. Mixed methods research is the type of research in which a researcher or team of researchers combine elements of qualitative and quantitative research approaches (e. g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration (Johnson et al. 2007, p. 123). Mixed Methods-Design (Quan + qual) (Guest 2013) was adopted as qualitative data were used as dominate.

Data Collection Method

The participants of this research are mathematics teachers of 12 different elementary schools. The participant teachers were asked to evaluate six different students' problem solutions with general evaluation, a rating scale, and a rubric. Then, the teachers' opinions about these three assessment methods were collected.

First, a mathematics problem was prepared in a question form and six different students were asked to solve these problems. The answers obtained from these six students were given to the teachers to be evaluated using three different methods.

The first method was general evaluation. The answers of the students were given to the mathematics teachers without specifying any scoring criteria, and they were asked to score through general evaluation. After general evaluations were taken, teachers were given the rating scale prepared by the researcher and were asked to evaluate students' answers based on this scale. Finally, the teachers were asked to evaluate the answers with the rubric prepared by the researcher

for these problems. With these three processes, six different students' problem-solving skills were evaluated based on general evaluation, the rating scale, and the rubric and the results were recorded. The data obtained were used to compare the different evaluation methods used by the teachers in terms of scoring tendencies and reliability.

One week after the application, interviews were conducted with the same raters to determine their views and the scoring strategies they applied in these three assessment methods. With the data obtained from the interview form, the opinions of the teachers on the evaluation methods and their preferences were identified.

Instruments

During the data collection process, the study used a rating scale, a rubric, and an interview form. There was no guidance in the general evaluation and teachers were asked to score according to their own assessment criteria.

Rating scale. There are four dimensions in the rating scale developed to evaluate problem-solving skills. They are as follows: Understanding the problem, determining a solution strategy, problem-solving, and checking the result of the problem. The teachers were asked to evaluate according to these dimensions and to give a score between 1 and 3.

Rubric. Rubrics were obtained by creating criteria for the dimensions used in the rating scale. When determining the criteria, possible student responses were taken into consideration. The teachers were asked to evaluate student responses considering these criteria.

Questionnaire. An interview form consisting of questions about which assessment tool teachers prefer to use and whether they tend to score high or low was developed and conducted with the participant teachers.

Analysis

Quantitative analysis – Many facet rasch analysis

FACETS software developed by Linacre (1994) was used for the analysis of the observation data obtained in the study. In this research, there are three surfaces to be analysed in the Rasch measurement model. The rater severity/leniency are four dimensions and 12 raters used to score skills. Since the aim of this study was to reveal the scoring tendencies of the raters and to determine their reliability, the reported findings were limited to reporting them. The data obtained from general evaluation, the rating scale, and the rubric were analysed separately, and the results were presented comparatively.

In Many Facet Rasch analysis, fit statistics, separation index, and separation index reliability (R) and chi-square statistics were calculated for each facet.

Fit statistics

This statistics show how much the expected scores match the observed scores (Engelhard & Myford, 2003; Linacre, 1989; Wright & Linacre 1994). In other words, it gives information about the degree of fit of the data to the measurement model (Lee & Kantor, 2003). If the data fit the model, researchers can provide useful and informative comparisons between sources of variability (Engelhard & Myford, 2003). Large differences between observed and expected scores (expressed as standardized residues) are indicative of surprising or unexpected results. These residual values are shown as the mean of error squares statistics called outfit and infit. Outfit statistics gives an index of the average weighted squares of residual values between expected and observed points (Engelhard, 1994). Outfit statistics are very sensitive to unexpected endpoints. Infit statistics weighted residual values are the mean squares. It is less sensitive to unexpected endpoints due to residual value statistics (Engelhard & Myford, 2003; Iramaneerat et al., 2008).

These statistics have the same distribution and interpretation, and the acceptable range of these statistics is between 0.8 and 1.2 (Linacre, 1994). The values obtained in this range can be described as efficient, thus, it is possible to conclude that the data model is fit.

Separation index and separation index reliability

Both indexes provide information about the reliability value. Separation index provides information on the degree to which all elements of each variance source are separated from each other (Lee & Kantor, 2003). In other words, it gives a measure of the spread (variability) of the precision of the source of variability. The second reliability value is separation index reliability (R). This index provides information on how well the elements in a source of variability can be reliably separated to identify the source of variability. This index is similar to traditional reliability statistics such as KR-20 Cronbachs' Alfa (Bond & Fox, 2001; Engelhard & Myford, 2003; Myford & Wolfe, 2003; Sudweeks et al., 2004).

Separation index reliability for each variance source was 0.0 to 1.0; the index of separation ranges from 1 to infinity (Sudweeks et al., 2004). The fact that the separation index reliability is close to 1.0 is indicative of a high level of reliability and is a desirable level (Bond & Fox, 2001). For the student variability source, the separation index and the separation index reliability are desirable to have a high value whereas, for other variability sources, it is desirable to have a low value because variability between the elements in other sources of variability is an indicator of an undesirable variance in the scores (Engelhard & Myford, 2003; Sudweeks et al., 2004). This index gives the spread of rater severity levels. An index of 1.0 can be considered as an indicator that raters score at similar severity levels and may be interchangeable, and this situation is desirable for raters (Engelhard & Myford, 2003). Low values of these two statistics can be interpreted as the measurements obtained for different elements of the source of variability as it shows a high degree of stability (no inconsistency) (Sudweeks et al., 2004).

Chi-square statistics

Chi-square statistics are used to calculate whether there is a significant difference between the sources of variability. In other words, the chi-square test is used to test whether there is a significant difference between the severity levels of the raters. A significant chi-square statistic ($p < .05$) indicates that there is a difference between at least two of the rater severity levels (Myford & Wolfe, 2004).

Qualitative data analysis

The responses of the participants to the questions in the interview form were analysed using content analysis. Content analysis was performed by two different encoders. First of all, coding rules were determined, and the encoders obtained the categories and the themes by coding the data separately. The reliability coefficient based on inter-coders compliance (Miles & Huberman, 1994) was recorded as .87.

Qualitative and quantitative data analyses were performed separately, and reported findings were interpreted together. The findings from the qualitative data analysis were used to follow through the findings from the quantitative data analysis.

Results

In order for the data used in the analysis to be compatible with the model, the absolute value of less than about 5% of the standardized values (*z* score) must be greater than or equal to 2, or, less than about 1% of the standardized values must be lower than or equal to 3 (Linacre, 2003).

Table 1. *Standard Values for Model Fit*

	Number of Observation	General Evaluation	Rating scale	Rubric
+/- 3	288	2 (.007 %)	1 (.003 %)	1 (.003%)
+/- 2	288	3 (.010 %)	3 (.010 %)	2 (.007 %)

When Table 1 is analysed, it can be seen that z values are in the required range. According to this model fit was provided for the main analysis. First of all, whether there was a difference between the scoring tendencies of the raters in all three assessment methods was tested. The hypothesis "There is a significant difference between raters in terms of their severity/leniency" was tested with the Chi-Square. In addition, the reliability of the raters was estimated for all three methods.

Table 2. *Mode Fit and Raters' Reliability*

	General Evaluation	Rating Scale	Rubric
RMSE (Model)	.24	.11	.15
Fixed (all the same) chi-square	146.4	139.8	144.9
d.f.	11	11	11
significance	.00	.00	.00
Random (normal) chi-square	14.6	13.1	13.9
d.f.	10	10	10
significance	.36	.26	.37
Separation index (for raters)	7.23	4.94	6.51
Reliability (for raters)	.84	.63	.79

When the separation indices presented in Table 2 are examined, it is seen that the smallest value is obtained from the situation where the rating scale was used. The highest separation index is obtained from the general evaluation condition. The high index of separation shows that the scores given by the raters are different. Accordingly, the highest differentiation in the scores given by the raters is in the general assessment condition. The smallest difference in the scores given by the raters is in conditions in the evaluations with the rating scale. Accordingly, the scoring tendencies of the raters are closer when the rating scale is used.

As for the separation index reliability of the raters', a result similar to the separation index can be found. This high index shows that the raters gave different scores. However, the raters were expected to give similar scores and be consistent with each other. The separation index reliability coefficients of the raters' were found to be .84 for the general assessment method, .63 for the rating scale, and .79 for the rubric. Accordingly, the scores given by the raters are more homogeneous than the other conditions when the rating scale was used. In other words, more consistent scores were given by the raters when the rating scale was used. The highest differentiation between the scores given by the raters was in the condition that the general evaluation method was used.

When the chi-square values presented in Table 2 are examined, it is seen that there is a significant difference between the leniency and the severity of the raters in all three evaluation methods ($\chi^2_{\text{general evaluation}}$ = 146.4 *sd* = 11, *p* = .00; $\chi^2_{\text{rating scale}}$ = 139,8 *sd* = 11, *p* = .00; χ^2_{rubric} = 144.9 *sd* = 11, *p* = .00). Therefore, the null hypothesis was rejected. Accordingly, there is a significant difference among the leniency/severity of the raters. Table 3 presents the raters' leniency/severity rankings.

Table 3. *The Raters' Leniency/Severity Rank*

	General evaluation			Rating scale			Rubric		
	Rater	Infit	Outfit	Rater	Infit	Outfit	Rater	Infit	Outfit
	11	1.2	1.1	11	1	1.1	11	1.3	1.2
	7	1.1	1.3	7	0.8	0.8	3	1.1	1.2
	4	1.18	1	8	0.8	0.9	7	1.1	1
	3	1.18	1	3	0.9	1	2	1.1	0.9
	6	1	1.1	2	1.1	1.1	6	0.9	0.9
	12	0.8	0.9	6	1.1	1	9	0.8	0.8
	1	0.8	0.9	1	1.1	1.1	12	0.8	0.9
	9	0.9	0.8	9	1.2	1	8	1	0.9
	8	0.9	0.8	5	0.8	0.9	10	0.8	0.7
	10	1.1	0.7	12	0.9	0.9	4	1.2	1.1
	5	1.1	0.7	4	0.9	1	1	1	1.1
	2	1.0	1,1	10	1.1	1.2	5	0.8	0.9

Severity

The quality control limit specified in the "infit" and "outfit" values in Rasch analysis is between 0.6 and 1.4 (Wright & Linacre, 1994, p. 375-380). Accordingly, the raters made appropriate scoring in all evaluation conditions. The most severe rater was rater 2 when the general evaluation method was used, it was rater 10 when the rating scale was used, and rater 5 when the rubric was used. The most generous rater was rater 11 in all scoring methods. In Figure 1, the raters' leniency/severity rankings between -1.25 logit and 2.25 logit are presented.

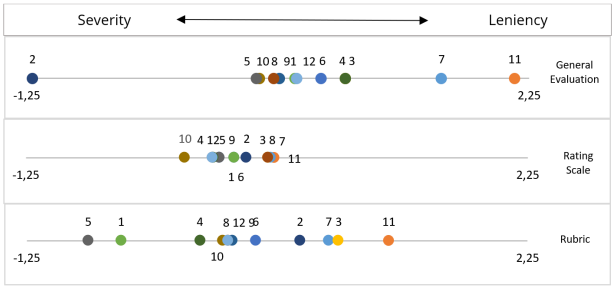

Figure 1. The Raters' Leniency/Severity Rank

When the general evaluation method is used, it is observed that the leniency/severity tendency of the raters are in a widest range, followed by the rubric. The rating scale case showed the smallest range among the three methods. Accordingly, the scoring tendencies of the raters became similar when the rating scale was used, while the raters' tendencies differed in the general assessment. Therefore, the use of the rating scale method enabled the raters to be objective at the highest degree. In the general evaluation method, the raters were found to be more subjective. In other words, it can be stated that the error rate caused by the rater factor was higher.

When the rating scale was used by the raters, teachers' leniency/severity tendencies approached each other. Reliability in terms of objectivity is provided by different raters giving similar scores. Objectivity is particularly affected by errors caused by the rater. When the rating scale is used, the raters' tendency to score is closer to each other when compared with the other two assessment methods, and it can be claimed that they make more objective scoring. General evaluation was the method most affected by the raters' personal judgments, and therefore the objectivity is the lowest. As the raters' tendencies obtained in the general evaluation are examined, it is

Table 4. *The Results for Differences in The Raters Scorings*

Assessment tools	Raters	Negative Rank	Positive Rank	Tieg	Negative Ranks Mean	Positive Rank Mean	Sum of Negative Ranks	Sum of Positive Ranks	*z*	*p*
General evaluation	2	0	4	2	0	2.5	0	10	-2.000	.046*
	3	4	0	2	2.5	0	10	0	-2.000	.046*
	7	5	0	1	3	0	15	0	-2.070	.038*
	11	5	0	1	3	0	15	0	-2.041	.041*
Rating scale	7	5	0	1	3	0	15	0	-2.060	.039*
	11	5	0	1	3	0	15	0	-2.060	.039*
Rubric	3	6	0	0	3.5	0	21	0	-2.201	.028*
	5	0	6	0	0	3.5	0	21	-2.214	.027*
	7	5	0	1	3	0	15	0	-2.060	.039*
		6	0	0	3.5	0	21	0	-2.220	.026*

*$p < .05$

seen that the range is the largest. When rubric was used as the evaluation method, a more homogeneous result than the general evaluation. Thus, it can be stated that the most objective method is the rating scale, and the least objective method is the general evaluation method.

The responses of six different students were evaluated by the professors in five different mathematics fields using general evaluation, the rating scale, and the rubric. The student scores determined by these five different evaluators and the scores given by the raters were compared in pairs with the Wilcoxon test. Whether there is a significant difference between the raters' scores and the scores determined by the professors was tested. Test results are presented in Table 4.

When Table 4 is examined, it is revealed that some of the raters scored higher or lower than expected. Four raters in the general evaluation method and four raters in the rubric method gave significantly higher or lower scores than they should have whereas in the rating scale method, two raters scored significantly higher or lower than expected. Furthermore, rater 11 and rater 7 had the tendency to give high scores in all of the assessment tools used, in other words, their tendency of giving score did not depend on the assessment tools. It is seen that rater 2 tended to give low scores when the general evaluation was done. However, the rating scale and the rubric were used, the rater changed the tendency to give low scores and gave accurate scores. Similarly, rater 5 gave low scores when the rubric was used to assess, but in the other cases, the rater gave accurate scores. So it can be claimed that rater 5 lost the tendency to give objective scores when the rubric was used. In the case of general evaluation and the rubric rating, more raters gave significantly higher or lower scores than in the case of rating scale. Therefore it can be said that when the raters used rating scales to assess the students' answers, they were more objective than the other cases where general evaluation or evaluation with the rubric were used.

The results of the content analysis of the raters' opinions about the type of the assessment tools and their own tendencies to give scores are presented below. Table 5 presents the raters' opinions on their scoring tendencies.

Table 5. *Raters' Views on Their Own Scoring Behaviour*

	Scoring tendencies		
	High scores	Objective scores	Low scores
Raters	3*, 5*, 6, 7*, 8, 10, 11*	1, 2*, 4, 9,1 2	-

Table 5 presents the evaluators' scoring tendencies. Accordingly, rater 3, 7, and 11 had the tendency to give score significantly higher. The opinions of these raters confirm their tendency to give high scores. In other words, these raters were aware of their tendencies. Raters 2 and 5 tended to give low scores in some cases, however, in the interviews, rater 5 claimed to give higher scores while rater 2 stated to score objectively. Considering that they did not exhibit the same tendencies in all conditions, it can be stated that there is a tendency to be exhibited depending on the evaluation method used.

Table 6. *Approaches of Raters for Assessment Tools*

	General evaluation	Rating scale	Rubric
Prefer to use	11, 8	1, 5, 9, 12	2, 3, 4, 6, 7, 10
Difficult to use	1, 3, 4, 5, 6, 7, 10, 11	2, 9	8, 10, 11, 12

Only two of the raters preferred the general evaluation method while the other raters indicated that they preferred rubrics. While the majority of the raters considered the general evaluation method as difficult, the rating scale was evaluated as difficult by two raters. Table 7 provides explanations of why raters find assessment methods difficult. The raters' opinions were classified in psychometric property, evaluation characteristics, and application themes. The explanations of the themes are presented preceding the tables:

Psychometric property

It includes phrases about the validity and the reliability of assessment tools. For example, the term `consistent` is related to reliability, so if a rater mentions it, it is classified under this theme.

Evaluation characteristics

It includes phases of assessment tools such as descriptions of tasks, criteria for evaluation of answers.

Application

It includes a statement about the application of assessment tools. For example, if the rater says that it takes a long time to develop or implement an assessment tool, this opinion is classified under the application theme.

It is seen that lack of criteria, being subjective, and giving points by comparing students were expressed as negative

Table 7. *The Raters' Views on The Negative Aspects of The Assessment Tools*

	General evaluation		Rating scale		Rubric	
Theme	Categories	Rater	Categories	Rater	Categories	Rater
Psychometric property	Subjective	1, 5, 6, 10, 12				
Evaluation characteristics	Without criteria	7, 11, 1, 3, 4, 5, 6	Descriptions are not enough	10, 2		
	Requires benchmarking.	3, 5, 10	Not enough to show differences within students' levels.	2		
Application					Take more time for developing and applying.	9, 10, 11, 12
					Difficult to understand	12

Table 8. *The Raters' Views on The Positive Aspects of The Assessment Tools*

	General evaluation		Rating scale		Rubric	
Theme	Categories	Rater	Categories	Rater	Categories	Rater
Psychometric property			It is consistent	5	It is consistent	3, 8, 10
Evaluation characteristics			Criteria are understandable.	1	Criteria and explanations are clear and understandable.	6, 7, 3, 4
			Includes enough explanation.	12, 1	The process to be followed is systematic and clear.	4, 7
			Considers all necessary skills	9	Includes all answer categories	10, 2
Application	Allows creating your own criteria.	11, 8	Saves time	1,9,12		

characteristics of general evaluation method. However, general evaluation method was not criticized in terms of application. The lack of an adequate definition of the rating scale and not enabling showing differences within students' levels were categorized as negative characteristics. It was not criticized in terms of psychometric properties or practice. For rubric, the negative features were difficult to understand and difficult to develop and implement. It was found that the characteristics of psychometric and method characteristics were not criticized. Accordingly, rubrics are found to be difficult only in terms of usability.

The opinions about the positive aspects of the instruments are presented in Table 8.

When Table 8 is analysed, it is seen that the general evaluation method is preferred because of only one reason, which is allowing creating your own criteria, and just two raters supported the method for this reason. Therefore, it can be said that the other raters thought that this method is not suitable for assessing student achievement. In other words, the raters did not think that the general evaluation method has any psychometric property or excellent characteristics property. Four raters preferred to use rating scales when assessing students' achievement because of characteristics aspect such as having clear/understandable criteria, including sufficient explanation for criteria and considering necessary skills for assessment. In addition, one rater stated to prefer psychometric property of the rating scale since it is consistent. Six raters who preferred to use rubrics expressed that rubrics have specific characteristics such as criteria, providing clear/understandable explanations, enabling to follow a systematic and a clear process, and including all categories of students' answers. In addition, three raters thought that rubrics are consistent. No raters mentioned that they preferred rubrics because it has excellent application property.

Discussion and Future Directions

In this study, it is aimed to examine the scoring tendencies of the raters depending on the assessment method used and to determine which assessment method they prefer and why. Also, the scoring reliability of the raters was estimated with the separation index and the separation index reliability. The analysis was done based on IRT Many Facet Rasch Model. The reliability of the separation index was estimated as .84 for the general evaluation, .79 for the rubric, and .63 for the rating scale. The separation index was estimated as 6.23 in the general evaluation condition, 5.51 in the rubric condition, and 3.94 in the condition where the rating scale was used. High separation index reliability and separation index are desirable for students' level or scoring criteria. However, the low separation index of raters means that the raters have similar scoring tendencies; therefore, it is desirable to have a low separation index and separation index reliability when the raters are in question (McNamara, 1996; Myford & Wolfe, 2004). Accordingly, when the methods are compared, it can be stated that when the general assessment method is used, the raters' behaviours differ, and when the rating scale is used, the raters give similar scorings to each other. Therefore, the most objective evaluation is found to be in the condition of the rating scale.

A significant difference was found between the scoring tendencies of the raters in all three methods. Four of the raters scored significantly lower or higher than the required values when using the general evaluation method. Similarly, four of the raters scored higher or lower when using the rubric method. Only two of the raters scored higher when using the rating scale. It is also found that the number of raters showing significant differences is consistent with the separation index, and separation index reliability of the methods. Raters' tendencies affect the reliability of scorings (Black, 1998). Accordingly, it

is expected that raters that score higher or lower than required should have a lowering effect on scoring reliability. Accordingly, it can be claimed that in the case of using the rating scale, the raters scored more consistently and similarly compared to the other methods. Therefore, the reliability of the scoring with the rating scale is higher. Some studies in the literature also reveal that scoring tendencies of raters may differ and that rater reliability is influenced by rater behaviours such as leniency and severity (Güler, 2014; Brookhart, Walsh, & Zientarski, 2006; Mulqueen, Baker, & Dismuskes, 2000).

Although scoring tendencies of raters cannot be controlled, the reliability of the scoring can be increased by the evaluation method chosen. Giving criteria to raters for the scoring process affects their assessment (Eckes, 2008; Li & Lindsey, 2015; Schaefer, 2008; Tan & Turner, 2015). The fact that no criteria were given during the general evaluation and the different behaviours of the raters under the general evaluation condition support this opinion. The absence of any criteria during the general evaluation required the raters to form their own criteria. As each rater's criteria can be different, their scoring will be different as well. Therefore, the condition with the highest index of separation was the condition where the general evaluation was done.

Rubric tells both teachers and students what is important and what to consider when evaluating. (Arter & McTighe, 2001; Busching, 1998; Perlman, 2003). Therefore, rubrics are the best way to assess complex competencies without compromising reliability and validity (Morrison & Ross, 1998; Wiggins, 1998). Considering that giving criteria to raters will affect evaluation process positively (Eckes, 2008; Eckes, 2012; Li & Lindsey, 2015; Tan & Turner, 2015), it is expected that scoring will be objective when using a rubric. However, in this study, the most objective scoring was obtained when the rating scale was used. It is stated that even if criteria are presented to raters, their tendencies may continue to affect the evaluation process (Cooksey, Freebody, & Wyatt-Smith, 200; Schaefer, 2008). Davidson, Howell, and Hoekema (2000) state that one of the most important reasons why two raters give different scores when it comes to rubric is experience difference. More heterogeneous rater behaviours can be argued to correlate with experience in rating. The difference between raters due to lack of experience cannot be eliminated completely but it can be minimized through trainings on using and developing rubric (Stuhlmann, Daniel, Dellinger, Denny, & Powers, 1999; Weigle, 1999). Rubrics are one of the most commonly used methods for developing and measuring mathematical skills (Shepard, 1989; Wilson, 1993; Anderson & Puckett, 2003; Docktor & Heller, 2009; Gadanidis, 2003; Moskal & Leydens, 2000; Szetela & Nicol, 1992). In order to make the right decisions on mathematical skills of students, measurement processes should be free of errors. Therefore, it should be known that errors caused by raters may be effective in the evaluation of performance tasks or open-ended items. It is argued that using rubrics is the best method in evaluation, so teachers' skills in developing and using rubrics are crucial in case of decisions about students (Romagnano, 2001).

The majority of the raters stated that they preferred rubrics in scoring, but it can be thought that the fact that four raters scored significantly higher or lower than required could be related to their experience with rubric use because the raters in the interviews stated that rubrics are difficult to understand and requires a lot of time while using and developing it. Not having enough experience may make understanding of the explanations in a rubric more difficult (Busching, 1998; Perlman, 2003; Wiggins, 1998). In addition, there are studies showing that although when the standards and criteria of the rubrics are clear, rating scales may be more reliable (Myford, Johnson, Wilkins, Persky, & Michaels, 1996; Penny, Johnson, & Gordon, 2000). In this study, it was also concluded that the most reliable method is the rating scale. The common characteristics are that both rubrics and rating scales include the basic criteria for evaluation. Rubrics also provide explanations to improve objectivity. However, although the raters knew the advantages of rubrics, unity was not achieved at the point of application about objectivity. As a result, their scores with the rubric were observed to be different in the study compared to other studies in the field. It can be claimed that the criteria in the rating scale were perceived by the raters in a similar way. For this reason, more objective evaluations were made when using the rating scale.

Ideally, an assessment should be independent of who does the scoring and the results need to be similar no matter when and where the assessment is carried out, but this is hardly obtainable. Although some traditional item types, for example with multiple-choice questions, meet more rigorous demands and are considered to be reliable, they are criticised for being insufficient in assessing complex performance. The more consistent the scores over different raters and occasions are, the more reliable the assessment is thought to be (Moskal & Leydens, 2000). Giving criteria in the assessment method was partly effective for objectivity. This is why the rating scale provided more objective results than the other two methods in the study. However, although the assessment method was changed, some raters continued to exhibit the same response behaviours. For example, rater 7 and rater 11 gave high scores in all circumstances. Schaefer (2008) states that the characteristics of the raters will affect the scoring regardless of the evaluation criteria. Similarly, Seker (2018) stated that factors such as education level, age, professional experience, and gender can be effective in rater behaviours. It can be argued that personal characteristics are the reason why some raters always have specific tendencies. For this reason, choosing an assessment method with certain criteria and referring to the opinion of more than one rater will provide more objective results.

In the study, problem-solving skills were used as a stimulus when examining rater tendencies. Future research can focus on examining raters' tendencies in assessing different skills and in evaluating different assessment methods such as performance tasks. The findings of the present study reveal that most of the participants prefer to use rubrics and are aware of the advantages of rubrics. The raters' scoring behaviours can be examined following trainings on the use of rubrics that provide them with activities to enhance their experiences. Thus, when rubrics are used for assessment, it can be demonstrated how having an experience of using rubric will affect scoring behaviours. In this study, Many Facet Rasch Model was preferred because individual evaluations were recorded and the raters' leniency/severity were determined. This model is not intended for generalization. A similar study involving a higher number of raters can be conducted grounded on a theory such as Generalizability theory.

Acknowledgements

This study was conducted under the supervision of Prof. Dr. Douglas McDougall (University of Toronto, OISE). I thank him for his contribution.

References

Aiken, L.R. (1996). *Rating scales and checklists: Evaluating behaviors, personality, and attitudes.* New York: John Wiley & Sons

Akın, Ö. & Baştürk, R. (2012). Keman eğitiminde temel becerilerin Rasch ölçme modeli ile değerlendirilmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 31*(31), 175-187.

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

Alharby, E.R. (2006). *A comparison between two scoring methods, holistic vs. analytic using two measurement models, the generalizability theory and the many facet rasch measurement within the context of performance asssessment.* (Unpublished doctoral thesis). The Pennsylvenia State University, USA

Anastasi, A. & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, NJ.: Prentice Hall.

Anderson, R.S. & Puckett, J.B. (2003). Assessing students' problem-solving assignments. *New Directions for Teaching and Learning*, (95), 81-87. http://dx.doi.org/10.1002/tl.117

Arter, J. & McTighe, J. (2001). *Scoring rubrics in the Classroom: using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press, Inc.

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

Baştürk, R. (2010). Bilimsel araştırma ödevlerinin çok yüzeyli Rasch ölçme modeli ile değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(1), 51-57.

Black, P. (1998). *Testing: Friend or Foe?* London: Falmer Press.

Bond, T.G. & Fox, C.M. (2001). *Applying the Rasch model: Fundamental Measurement in the Human Sciences.* London: Lawrence Erlbaum Associates.

Boztunç Öztürk, N., Şahin, M.G., & İlhan, M., (2019). An analysis of scoring via analytic rubric and general impression in peer assessment. *Turkish Journal of Education, 8*(4), 258−275.

Brookhart, S.M. & Walsh, J.M., Zientarski, W.A. (2006). The dynamics of motivation and effort for classroom assessment in middle school science and social studies. *Applied Measurement in Education, 19*(2), 151-184. http://dx.doi.org/10.1207/s15324818ame1902_5

Busching, B. (1998). Grading inquiry projects. *New Directions for Teaching and Learning, 74*, 89-96.

Casabianca, J.M. & Junker, B. (2013). *Hierarchical rater models for longitudinal assessments.* Paper in Annual Meeting of the National Council for Measurement in Education'. San Francisco, California.

Casabianca, J.M. & Junker, B. (2014). *The hierarchical rater model for evaluating changes in traits over time.* Paper in 121st Annual Convention of the American Psychological Association, Division 5: Evaluation, Measurement and Statistics, Washington D.C.

Çetin, B., & Kelecioğlu, H. (2004). The relation between scores predicted from structured features of essay and scores based on scoring key and overall impression in essay type examinations. *Hacettepe University Journal of Education, 26*, 19–26.

Çıkrıkçı, N. (2010). Üst düzey düşünme becerilerinin ölçülmesinde gündelik yaşam unsuru. *Cito Eğitim: Kuram ve Uygulama. 1*, 9-26.

Cooksey, R. W., Freebody, P. & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analyzing how teachers evaluate students'writing. *Educational Research and Evaluation, 13*(5), 401 434.https://doi.org/10.1080/13803610701728311.

Cooper, W. H. (1981). *Unbiquitous halo. Psychological Bulletin, 90*(2), 218-244.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* Harcourt Brace Javanovich College Publishers, USA.

Davidson, M., Howell, K. W. & Hoekema, P. (2000). Effects of ethnicity and violent content on rubric scores in writing samples. *Journal of Educational Research, 93*, 367–373.

DeCarlo, L.T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42*(1), 53-76.

DeCarlo, L.T. (2010). Studies of a Latent Class Signal Detection Model for Constructed Response Scoring II: Incomplete and Hierarchical Designs. ETS Research Report Series, (08). Princeton, NJ: Educational Testing Service.

DeCarlo, L.T., Kim, Y.K. & Johnson, M.S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*(3), 333-356.

Docktor, J. & Heller, K. (2009). *Assessment of student problem solving processes.* In AIP Conference Proceedings, 1179, 133-136. http://dx.doi.org/10.1063/1.3266696

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

Donoghue, J.R. & Hombo, C.M. (2000). *A comparison of different model assumptions about rater effects.* In Annual Meeting ofthe National Council on Measurement in Education Proceedings. New Orleans, LA.

Eckes, T. (2008). Rater types in writing performance assessments: A classification ap-proach to rater variability. *Language Testing, 25*(2), 155–185.https://doi.org/10.1177/0265532207086780.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition torater behavior. *Language Assessment Quarterly, 9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

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, (01). Princeton, NJ: Educational Testing Service.

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

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.

Gadanidis, G. (2003). Tests as performance assessments and marking schemes as rubrics. *Reflections, 28*(2), 35-40.

Güler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research, 55*, 73-90.

Haladyna, T.M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. USA: A Pearson Education Company.

Iramaneerat, C., Myford, C.M., Yudkowsky, R. & Lowenstein, T. (2009). Evaluating the effectiveness of rating instruments for a communication skills assessment of medical residents. *Advances İn Health Sciences Education, 14*(4), 575-594.

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 İn Health Sciences Education, 13*(4), 479-493.

Johnson, B.R., Onwuegbuzie A.J. & Turner, L.A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research. 1*, 112–133. doi: 10.1177/1558689806298224.

Junker, B.W. & Patz, R.J. (1998). *The hierarchical rater model for rated test items.* In Annual North American Meeting of the Psychometric Society Proceeding. Champaign-Urbana, IL.

Kastner, M. & Stangla, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia-Social and Behavioral Sciences, 12*, 263-273.

Kim, Y.K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model* (Unpublished Doctorial Thesis). Teachers College, Columbia University.

Lee, Y.W. & Kantor, R. (2003). *Investigating differential rater functioning for academic writing samples: an MFRM approach.* In Annual Meeting of National Council on Measurement in Education proceeding. Chicago, IL.

Li, J. & Lindsey, P. (2015). Understanding variations between student and teacher ap-plication of rubrics. *Assessing Writing, 26*, 67–79. https://doi.org/10.1016/j.asw.2015.07.003.

Linacre, J. M. & Wright, B. D. (2004). Construction of measures from many-facet data. In E.V. Smith ve R.M. Smith (Eds.), *Introduction to Rasch Measurement* (pp.296-321). Maple Grove, MN: JAM Press

Linacre, J.M. (1989). *Many-facet Rasch measurement* (Unpublished Doctorial Thesis). University of Chicago, USA.

Linacre, J.M. (1990). *A Facet Model for Judmental Scoring.* MESA Memo 61.

Linacre, J.M. (1994). *Many-Facet Rasch Measurement.* Chicago: MESA.

Linacre, J.M. (2003). The hierarchical rater model from a Rasch perspective. *Rasch Measurement Transactions (Transactions of the Rasch Measurement SIG American Educational Research Association), 17*(2), 928.

Linacre, J.M., Wright B.D. & Lunz M.E. (1990). *A Facets Model of Judgmental Scoring.* Memo 61. MESA Psychometric Laboratory. University of Chicago. www.rasch.org/memo61.html.

Lunz, M. E. & Schumacker, R. (1997). Scoring and analysis of performance examinations: a comparison of methods and interpretations. *Journal of Outcome Measurement, 1*(3), 219-238.

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. *Langauge Testing, 15*, 158-80.

Mariano, L.T. (2002). *Information accumulation, model selection and rater behavior in constructed response student assessments* (Unpublished doctorial thesis). Carnegie Mellon University, Pennsylvania

McNamara, T.F. (1996). *Measuring Second Language Performance.* London and New York: Longman.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Miles, MB. & Huberman, AM. (1994). *Qualitative Data Analysis* (2nd edition). Thousand Oaks, CA: Sage Publications.

Morrison, G. R. & Ross, S. M. (1998). Evaluating technology-based processes and products. *New Directions for Teaching and Learning, 74*, 69–77.

Moskal, B.M. & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10). Retrieved from http://areonline.net/getvn.asp?v=7&n=10

Mulqueen C., Baker D. & Dismukes, R.K. (2000). *Using multifacet Rasch analysis to examine the effectiveness of rater training.* Presented at the 15th Annual Conference for the Society for Industrial and Organizational Psychology (SIOP). New Orleans.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., Johnson, E., Wilkins, R., Persky, H. & Michaels, M. (1996). *Constructing scoring rubrics: Using "facets" to study design features of descriptive rating scales.* In Paper presented at the annual meeting of the American Educational Research Association.

Nakamura, Y. (2000). Many facet rasch based analsis of communıcative 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.

Ömür, S. ve Erkuş, A. (2013). Dereceli puanlama anahtarıyla, genel izlenimle ve ikili karşılaştırmalar yöntemiyle yapılan değerlendirmelerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 28*(2), 308-320.

Ozbasi, D. & Arcagok, S. (2019). An Investigation of Pre-Service Preschool Teachers' Projects Using The Many-Facet Rasch Model. *International Journal of Progressive Education, 15*(4), 157-173.

Park, Y. (2017). Examining South Korea's Elementary Physical Education Performance Assessment Using Assessment Literacy Perspectives. *International Electronic Journal of Elemantary Education, 10*(2), 201-213. https://doi.org/10.26822/iejee.2017236116.

Patz R.J., Junker B.W. & Johnson M.S. (2000). *The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data.* Revised AERA Paper.

Patz, R.J., Junker, B.W., Johnson, M.S. & Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*(4), 341- 384

Penny, J., Johnson, R.L. & Gordon, B. (2000). Using rating augmentation to expand the scale of an analytic rubric. *The Journal of Experimental Education, 68*(3), 269-287.

Perlman, C.C. (2003). *Performance Assessment: Designing Appropriate Performance Tasks and Scoring Rubrics.* North Carolina, USA.

Pollack, J.M., Rock, D.A. & Jenkins, F. (1992). *Advantages and disadvantages of constructed-response item formats in large-scale surveys.* Paper in annual meeting of the American Educational Research Association. San Francisco, California.

Popham, W.J. (2008). *Classroom Assessment What Teachers Need to Know.* USA: Pearson Education

Rodriquez, M. C. (2002). Choosing An Item Format. Tindal, G. ve Haladyna, T.M. (Ed.). *Large-Scale Assessment Programs For All Students* (213-231). New Jersey: Lawrence Erlbaum Associates Publishers.

Roid, G.H. & Haladyna T.M. (1982). *A Technology for Test-Item Writing.* New York: Academic Pres.

Romagnano, L. (2001). The Myth of Objectivity in Mathematics Assessment. *Mathematics Teacher, 94*(1), 31-37. Retrieved from http://www.peterliljedahl.com/wp-content/uploads/Myth-of-Objectivity.pdf

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Sebok, S. (2010). *"Pick me, pick me, ı want to be a counsellor" assessment of med. counselling application selection process using Rasch analysis and generalizability theory* (Unpublished master thesis). University of Northern British Columbia: USA.

Seker, M. (2018). Intervention in teachers'differential scoring judgments in assessing L2 writing through communities of assessment practice. *Studies in Educational Evaluation, 59*, 209-217. https://doi.org/10.1016/j.stueduc.2018.08.003.

Shepard, L.A. (1989). Why we need better assessments. *Educational Leadership, 46*(7).

Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R. K. & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Journal of Reading Psychology, 20*, 107–127.

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

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.

Szetela, W. & Nicol, C. (1992). Evaluating problem solving in mathematics. *Educational Leadership, 49*(8), 42-45. Retrieved from http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_199205_szetala.pdf.

Tan, M. & Turner, C. E. (2015). The impact of communication and collaboration betweentest developers and teachers on a high-stakes ESL exam: Aligning external assessmentand classroom practices. *Language Assessment Quarterly, 12*, 29–49. https://doi.org/10.1080/15434303.2014.1003301.

Verhelst, N. & Verstralen, H. (2001). IRT Models for Multiple Raters. In A. Boomsma, T. Snijders, and M. van Duijn, (Ed.), *Essays in Item Response Modeling*. New York: Springer-Verlag.

Wang, Z.G. (2012). *On the use of covariates in a latent class signal detection model, with applications to constructed response scoring* (Unpublished doctoral thesis). Columbia University, New York

Weigle, S.C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

Wiggins, G. (1998). *Educative Assessment.* San Francisco: Jossey-Bass.

Wilson, M. & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26*, 283–306.

Wilson, L. D. (1993). *Assessment in a secondary mathematics classroom.* (Ph.D. diss.), University of Wisconsin-Madison.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 8*(3), 370.

Wright, B. D. & Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press.