SAKARYA UNIVERSITY JOURNAL OF EDUCATION

Original Research Doi: 10.19126/suje.963046

Received: 06.03.2021 Accepted: 20.08.2021

Published: 30.08.2021 August 2021• 11(2) • 357-378

Group Moderation Assessment Model: An Example of an Open-Ended Mathematics Exam[*]

Mithat TAKUNYACI[**] Emin AYDIN[***]

Abstract. The purpose of this study is to examine the assessment of an open-ended mathematics exam to reveal the effects of the Group Moderation Assessment Model. The Group Moderation Assessment Model is a process in which teachers share their expectations and understanding standards with each other to improve the consistency of their students' learning decisions. In this study, one of the non-random sampling methods, the appropriate sampling method was used. The exam papers used in this study belong to a total of 22 students studying in the 10th grade. The students' exam papers (for three math exams) were evaluated by an assessment team of five mathematics teachers in the group moderation assessment model. The findings show that the raters were positively influenced by each other and that they formed a reliable evaluation system by making judgments. In addition, it was found that the raters scored in a consistent way with each other in the exams conducted after the group moderation assessment model workshops. In conclusion, in the workshops held during the implementation of the group moderation assessment model, it was found that the teachers' knowledge and opinion with each other positively affected the teachers' ability to assess exam papers.

Keywords: Open ended questions, reliability, group moderation, assessment, measurement and evaluation.

[*] Ethics committee approval for this study was obtained from Sakarya University Scientific Research and Publication Ethics Committee, dated 13/01/2021 and protocol number 30.

[**] Orcid ID: https://orcid.org/0000-0003-1065-975X, Assist. Prof. Dr., Sakarya University, Turkey, mtakunyaci@sakarya.edu.tr

[***] Orcid ID: https://orcid.org/0000-0003-4298-2623, Prof. Dr., Marmara University, Turkey, eaydin@marmara.edu.tr

Takunyacı, M., & Aydın, E. (2021). Group Moderation Assessment Model: An Example of an Open-Ended Mathematics Exam. *Sakarya University Journal of Education, 11*(2), 357-378. doi: https://doi.org/10.19126/suje.963046

1. INTRODUCTION

In recent years, competence in educational assessment has come to be seen as a basic skill for all teachers (De Luca & Johnson, 2017). Assessment helps focus each student's learning process and outcomes. The collection of student assessment information is necessary to develop teaching and learning strategies and to meet the information needs at the level of students, parents, teachers, school leaders, policy makers and the general public (Harlen 2005; Gipps & Stobart 2003; Maxwell & Gipps 1996; Shavelson, Yin, Furtak, Ruiz-Primo & Ayala, 2008; Strachan 2002).

In the literature, studies reveal a strong impact of different types of assessment on student learning outcomes (Hutchinson & Hayward, 2005; Malone, Long and De Lucchi, 2004; Maxwell 2007; Wilson 2004). Evidence on different approaches indicates that assessment may support or diminish student motivation and performance depending on the way it is designed, implemented and used. In other words, assessments that are not well designed and implemented may in fact contribute to alienating students (and teachers) from the education system and exacerbate inequity in education (Cooksey, Freebody & Wyatt-Smith, 2007). On the other hand, carefully planned assessment interventions that are well aligned with learning goals and place students at the center of the process have strong potential to raise achievement and reduce disparities. Students need to be clear about what they are aiming to learn, and which indicators and criteria are appropriate to evaluate progress and inform future learning (Klenowski & Wyatt-Smith, 2013; Sadler, 1998). Engaging students as active participants in assessment will help them develop capabilities in analyzing their own learning and becoming self-directed learners. Teachers need assessment information that is reliable and consistent across schools in order to understand student strengths and weaknesses in relation to expected standards, to target future teaching and improve classroom instruction (Harlen, 2005; Klenowski & Wyatt-Smith, 2010; Spiller, 2012). School leaders can use such information for school self-evaluation processes and to provide accountability information to their employers and the educational administration.

Measurement and assessment are carried out with the aim of monitoring the functioning of educational programs, the effectiveness of teaching methods and techniques, and determining student achievement and learning difficulties (Baykul, 2000). The most common purpose of the assessment process in schools is to reveal whether the expected behavior change has occurred and to answer the questions about the degree of achievement of the purpose of educational activities objectively (Özçelik, 1992). The right decisions to be made during the education and training process are related to the qualifications of the measurement tools. According to Turgut (1992), it is possible to make objective and accurate decisions in education by using sensitive measurement tools and methods.

Measurement is a frequently used tool in the evaluation of any educational process. Measuring instruments have similar and superior aspects to each other. It is not possible to measure when, how, where and why a student will apply what he knows and his competencies by performing multiple-choice tests (Stecher, 2010). For open-ended questions, it is not possible to score objectively. The scoring of the open-ended question may differ according to the person who scored. The biggest disadvantage of open-ended questions is that the reliability, which is one of the most important features of a measurement tool, is due to the rater, since it does not score objectively as in multiple-choice questions (Romagnano, 2001; Takunyacı, 2016).

Reliability as defined by Kerlinger (1992) is the accuracy or precision of a measuring instrument. Reliability testing determines if an instrument is free from error and provides consistent results (Association for Advanced Training, 1988; Humphry and Heldsinger, 2020). The more reliable a test, the less error there is in the test. Thus, the test can be expected to provide repeatable and consistent results (Association for Advanced Training, 1988).

According to Tekin (2000), for reliability of the open-ended exams, if possible two or more raters should read the answers separately. This method should definitely be used in exams especially for the purpose of identifying the awardees or selecting students for a further education program. Scores from only one rater for a test administered once are not completely reliable. The score an individual gets from different situations; different test forms or different raters often varies. The purpose of measurement studies is to obtain observed scores as close to real scores as possible. Measurement results close to real scores are realized in the extent of the low error scores in the measurements. In other words, measurement results are reliable to the extent that they have few random errors (Baykul, 2000; Humphry and Heldsinger, 2020).

According to Aiken (2000), the level of consistency of the ratings made by more than two raters for different items and individuals is defined as rater reliability. Inter-rater reliability is important when subjective opinions are needed to rate and score individuals, events, characteristics or behaviors (Goodwin, 2001). However, if rater reliability is not provided, a student's score may change from rater to rater, and students also state that their scores are generally based on the subjective judgment of the rater (Moskal and Leydens, 2000). When the literature is examined, it is seen that there are studies of rater reliability for open-ended or performance-based exams (Büyükkıdık and Anıl 2015; Doğan and Anadol, 2017; Güler and Gelbal 2010; Güler and Teker, 2015; Nalbantoğlu & Başusta, 2015; Kan, 2005). In the study conducted by Güler and Teker (2015), correlation, comparison of means, percentage of agreement and generalizability theory were used to determine inter-rater reliability and they stated that generalizability theory was the most useful method among these methods.

One of the frequently recommended methods in determining the scoring reliability is scoring the answers given to the test independently by different raters and examined the correlation between these scores (Gronlund & Linn, 1990). The fact that the correlation coefficient is close to 1 is interpreted that different rater score the answers given to the test similarly, in other words, the error made in the scoring is low (Büyüköztürk, 2014). The weak correlation or inconsistency of the scores obtained from different raters indicates that there is subjectivity in the scoring process, in other words, the scoring reliability is low. In such a case, the reliability can be increased by conducting more detailed studies about the source of the differences. If the number of raters is more than two, the scoring reliability will need to be examined in more detail (Aiken, 2000).

It has been observed that most of the studies on rater reliability in the literature have been carried out to determine the reliability of the tests applied in different countries (Evans-Hampton, Skinner, Henington, Sims & McDaniel, 2002; Güler & Gelbal, 2010; Stecker & Fuchs, 2000; Thurber, Shinn and Smolkowski, 2002). In addition, there are also studies that examined different methods for determining inter-rater reliability (Güler & Teker, 2015; Goodwin, 2001). However, no study has been found that handles the evaluation of open-ended exams, which have an important place in mathematics education, and rater reliability together.

Writing well-structured questions in open-ended exams alone is not enough for a quality exam process. Good scoring is as important as having good questions (Clarke, 2011; Downing, 2009; Reiner, Bothell, Sudweeks & Wood, 2002). Consistency between scoring is often used as a measure of good scoring. The presence of a high consistency between different raters who scored the same answer or the scoring made by the same rater at different times is expressed as evidence of scoring reliability (Shermis, Di Vesta, 2011). Ensuring the quality of the questions written does not guarantee high consistency in scoring. Scoring methods used to achieve this should also be appropriate.

Given the reliability, validity, and efficiency issues with assessing open-ended exams, group moderation assessment model (Bramley & Vitello, 2018; Wheadon, Barmby, Christodoulou & Henderson, 2019; Verhavert, Bouwer, Donche & De Maeyer, 2019) appears to have the potential to improve the quality of assessment (Benton & Gallacher, 2018; van Daal, Lesterhuis, Coertjens, Donche & De Maeyer, 2019). The group moderation assessment model is a concept developed in the United Kingdom to ensure comparability and consistency in the best possible way. Its main purpose is to improve the quality of assessment by increasing the reliability of open-ended exams (Gipps 1996; Gipps & Stobart 2003). Gipps (1994) stated that group moderation is to provide a scoring process that is not affected by the inconsistent grade of teachers or other external factors. The principal aim of group moderation in assessment was to improve the quality of assessment practice (Humphry & Heldsinger, 2019). Improvement of quality was seen as essentially a question of improving the validity of assessment tasks so that they would allow for the appraisal of important learning outcomes in each subject area (Allal & Mottier lopez, 2014; Black, Harrison, Hodgen, Marshall & Serret, 2010; Smaill, 2018; Smaill, 2020). In addition, improved validity was considered to be a way in which assessment would support student learning and thus link the formative and summative functions of assessment (Black & Wiliam. 2010). Furthermore, in group moderation assessment model, teachers share with each other their expectations and understanding of what learning looks like by the studied examples of different type and quality of students' work in order to improve the consistency of their students' decisions about their learning. This will help teachers to increase the reliability of the assessment information they collect (Allal & Mottier Lopez, 2014). This will enable teachers to make more reliable decisions about students' learning.

The problem statement of this research is to investigate the effect of the number of raters on the reliability of the scores in the group moderation assessment model in the evaluation of an open-ended mathematics exam process. This study aims to describe the effect of the number of raters and the reliability of the group moderation assessment model based on the results obtained from real data. This assessment model is expected to contribute to the field on inter-rater reliability.

2. METHOD

Research Model

This study is described as descriptive research which aim to examine the assessment of an open-ended mathematics exams to reveal the effects of the group moderation assessment model. In the group moderation assessment model, the reliability calculations made on the teachers' assessment about the exam papers constituted the quantitative part of the study.

Working group

In this study, one of the non-random sampling methods, the appropriate sampling method was used. In appropriate sampling, researchers select participants from geographically close, easily accessible, suitable and volunteered individuals (Gravetter & Forzano, 2012). The exam papers used in this study belong to a total of 22 students, 12 (54.5%) male and 10 (45.5%) female, studying in the 10th grade of a private high school located in Adapazarı district of Sakarya province. In addition, six mathematics teachers took part in the study as raters. One of the six mathematics teachers (male) included in our group adaptation assessment model was chosen as the lesson narrator, and the group moderation assessment model was studied with the other five (three male and two female) teachers.

Research Process

In the study, before the research, six high school mathematics teachers working in a foundation school in Adapazarı district of Sakarya were given a seminar by the researchers about the group assessment model. Later, these seminars were held at regular intervals (five times) throughout the semester. In these seminars, the points that the raters should pay attention to while evaluating the exam papers were discussed and common answer papers were prepared. The basic starting point of the group moderation assessment model applied in the assessment of students' exams is that the teachers conducting the course do not make the exam assessments of the students themselves. The course teacher attended the workshops only as an observer and was asked to assess the exam papers as an external rater. In this way, consistency and harmony in the assessment of teachers will be observed and, it is aimed, the bias in the assessment of the exam paper of the teacher giving the course, is eliminated. The implementation continued in the same way for the three mathematics exams throughout the term (Figure 1).

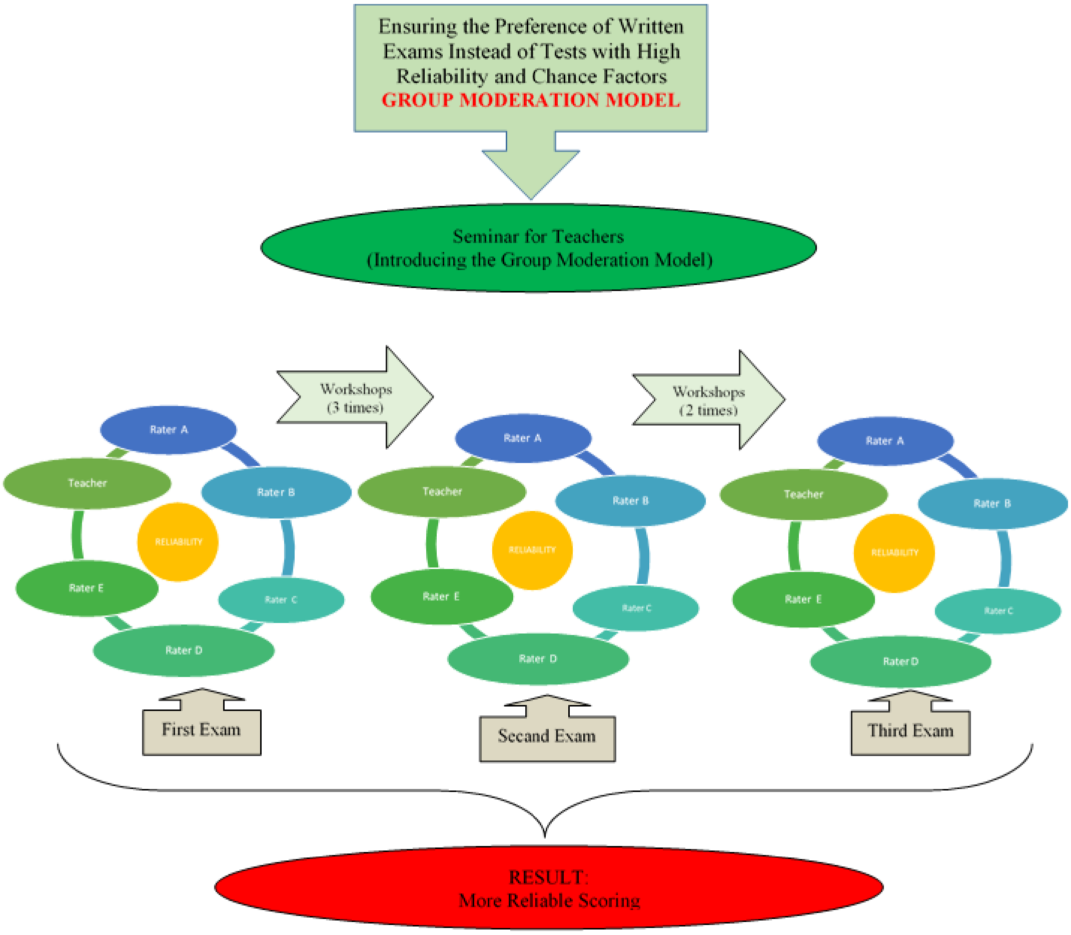


Figure 1. Research Process

3. RESULTS

As the first finding of the study, the reliability values were calculated for the internal consistency of the total scores given by the raters to the exams held before (first exam) and after (second exam, third exam) workshops. (Table 1).

Table 1

Cronbach's Alpha values for internal consistency between raters according to the total scores of the exams

	1st Exam	2nd Exam	3rd Exam
Raters	Cronbach's Alpha	Cronbach's Alpha	Cronbach's Alpha
Rater A	.861	.894	.946
Rater B	.813	.927	.968
Rater C	.804	.902	.935
Rater D	.853	.893	.966
Rater E	.857	.941	.955
Teacher	.805	.937	.960

When Table 1 is examined, it is seen that the internal consistency reliability values of the total scores given by all raters to the second and third exam papers have increased. It can be said that the workshops given within the scope of the group moderation assessment model had an effect on the raters' judgement of the exam papers consistently.

As the second finding of the study, it was tested with one-way analysis of variance whether there was a statistically significant difference between the reliability values calculated for the mean scores of the raters on the first, second and the third exam papers. The average reliability values of the scores given by the raters to the questions in each exam (10 questions for per exam) were calculated and 10 reliability values were calculated for each exam in total (Table 2).

Table 2

Average Cronbach's Alpha reliability values calculated by the raters.

	1st Exam		2nd Exam		3rd Exam	
Questions	average values	reliability	average values	reliability	average values	reliability
Q_1	.893*		.847*		.880*	
Q_2	.856*		.863*		.907*	
Q_3	.862*		.893*		.913*	
Q_4	.888*		.884*		.866*	

Q_5	.816*	.870*	.869*
Q_6	.864*	.901*	.904*
Q_7	.858*	.900*	.923*
Q_8	.862*	.898*	.918*
Q_9	.854*	.935*	.903*
Q_{10}	.806*	.883*	.915*

*p<.01

In the study, Shapiro-Wilks test was used for the normality test of 10 reliability values calculated for each exam. If the group size is less than 50, the Shapiro-Wilks test is used to examine the conformity of the scores to normality (Büyüköztürk, 2014). It was found that our data did not deviate excessively from the normal distribution since the p value calculated as a result of the normality test was higher than .05 (Table 3, p> .05).

Table 3

Shapiro-Wilks test result of Cronbach's alpha reliability values

	Shapiro-Wilks		
	Statistics	df	p
Average of Cronbach's alpha values for 1st Exam	.885	10	.149
Average of Cronbach's alpha values for 2nd Exam	.953	10	.699
Average of Cronbach's alpha values for 3nd Exam	.947	10	.630

The homogeneity distribution of the average reliability values of the scores given by the raters to the questions in each exam (10 questions for per exam) was made by looking at the Levene test. As a result of the analysis, it was found that the distribution of Cronbach's alpha reliability values was homogeneous (Table 4, p> .05).

Table 4

Levene Homogeneity Test results of Cronbach's alpha reliability values

	Levene Statistics	df1	df2	p
Average of Cronbach's alpha values for 1st Exam	.398	2	27	.676
Average of Cronbach's alpha values for 2nd Exam	1.351	2	27	.276
Average of Cronbach's alpha values for 3nd Exam	.331	2	27	.721

According to the results of Shapiro-Wilks normality test and Levene Homogeneity test, the data was found to be suitable for one-way analysis of variance for independent samples. In addition, the effect size values were also calculated in analyzing the research effects and used in the evaluation of the analysis of the findings. Cohen (1988), tried to classify the significance degrees of the effect size values in the model developed to facilitate the interpretation of effect size values. According to this classification, d ≤ .20 values, each of which are approximate values, are small, .20 <d <.80 values are medium and d ≥ .80 values are meaningful effect sizes.

Table 5

One-Way ANOVA test result for the average Cronbach's alpha reliability values calculated for the scores given by the raters on the first, second and third exam papers

	Sum of Squares	Sd	Mean Square	F	p
Between Groups	.010	2	.005	8.759	.001**
Within Groups	.016	27	.001		
Total	.026	29			

**p< .01

In Table 5, One-Way ANOVA test results are given for the average Cronbach's alpha reliability values of the points that the raters gave to each question in three written exams. According to these results [F = 8.759; p = .001 <.05], there is a statistically significant difference between the average Cronbach's alpha values calculated for the scores given by the raters on their written papers. The results in Table 6 were obtained when the Least Significant Difference (LSD) paired comparison test was applied to find out which written exams included the difference in question.

Table 6

Multiple Comparisons for Least Significant Difference (LSD)

(I) Exam	(J) Exam	Mean Difference (I-J)	Std. Error	p	Effect Size
1st	2nd	-.031*	.010	.007	.66
1st	3nd	-.043*	.010	.000	.91
2nd	3nd	-.012	.010	.262	.25

*p< .05

According to the LSD test results given in Table 5, although there is no statistically significant difference between the average Cronbach's alpha reliability values of the second and the third exams, between the first and second exam average Cronbach's alpha reliability values and between the first and third exam average Cronbach's alpha reliability values, A statistically significant difference was found in favor of the Cronbach's alpha reliability values of the second and third exams.

When we look at the effect size values, it can be said that the applied group model is moderately effective, since the effect size value obtained for the first and second exam average Cronbach's alpha reliability values is in the range of .20 <.66 <.80. Since the effect size value obtained for the first and third exam average Cronbach's alpha reliability values is .91 (> .80), it can be said that the applied group fit model is highly effective.

As the third finding of the study, the effect of the group moderation assessment model was examined according to the correlation values between the total scores given by the raters to the exam papers (Table 7, Table 8, Table 9).

Table 7

Binary correlation values of the scores given by the course teacher and raters to the exam papers for the first exam

Teacher	Rater A	Rater B	Rater C	Rater D	Rater E	
Teacher	1					
Rater A	.939**	1				
Rater B	.929**	.860**	1			
Rater C	.908**	.856**	.918**	1		
Rater D	.926**	.847**	.901**	.893**	1	
Rater E	.908**	.884**	.876**	.892**	.825**	1

** p< .01

As a result of the correlation analysis, it can be said that there is a statistically positive significant relationship between the first exam total scores of the raters. Looking at Table 7, it can be said that there is the highest correlation between the first exam score of the course teacher and the score of the A rater (r = .939), in other words, the scores of the course teacher and the A rater are quite close to each other. Among the scores, the lowest correlation is seen between the D and E raters (r = .825).

Tablo 8

Binary correlation values of the scores given by the course teacher and raters to the exam papers for the second exam

	Teacher	Rater A	Rater B	Rater C	Rater D	Rater E
Teacher	1					
Rater A	.958**	1				
Rater B	.946**	.917**	1			
Rater C	.919**	.893**	.937**	1		
Rater D	.926**	.861**	.938**	.919**	1	
Rater E	.925**	.886**	.942**	.937**	.939**	1

** p< .01

As a result of the correlation analysis, it can be said that there is a statistically positive significant correlation between the second exam total scores of the raters. Looking at Table 8, it can be said that there is the highest correlation (r = .958) between the second exam score of the course teacher and the score of the A rater, in other words, the scores of the course teacher and the A rater are quite close to each other. The lowest correlation between the scores is seen between the A and D raters (r = .861).

Tablo 9

Binary correlation values of the scores given by the course teacher and raters to the exam papers for the third exam

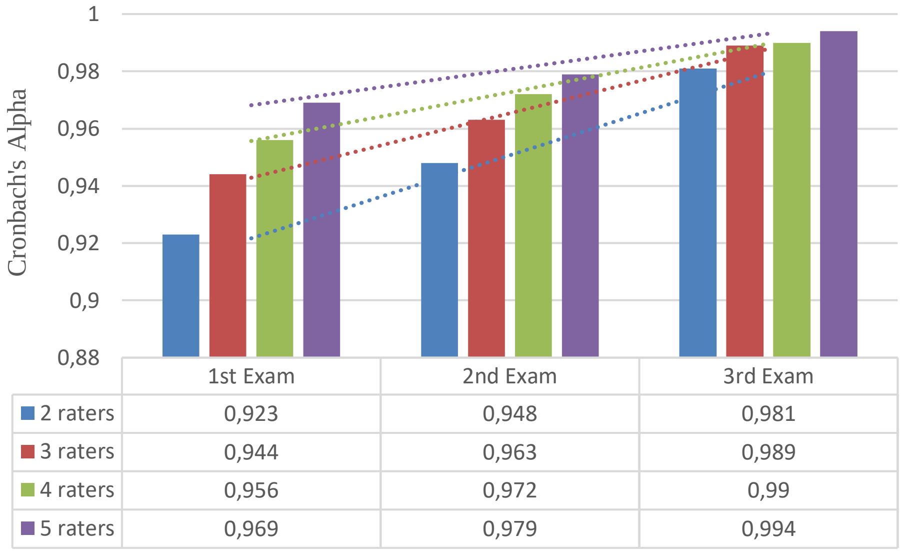
	Teacher	Rater A	Rater B	Rater C	Rater D	Rater E
Teacher	1					
Rater A	.988**	1				
Rater B	.986**	.978**	1			
Rater C	.972**	.971**	.986**	1		
Rater D	.981**	.966**	.973**	.953**	1	
Rater E	.980**	.977**	.970**	.967**	.981**	1

** p< .01

As a result of the correlation analysis, it can be said that there is a statistically positive significant correlation between the third exam total scores of the raters. Looking at Table 9, it can be said that there is the highest correlation coefficient between the third exam score of the course teacher and the score of the A rater (r = .988), in other words, the scores of the course teacher and the A rater are very close to each other. The lowest correlation between the scores is seen to be between the raters C and D (r = .953).

When the results of the correlation analysis were examined, it was observed that the correlation values between the scores of the raters increased statistically significantly in a positive direction. The correlation values for the first exam ranged between .825 and .939, the correlation values for the second exam ranged from .861 to .959, and the correlation values for the third exam ranged from .953 to .992. According to these findings, it is seen that the group moderation assessment model consistently creates a positive effect on teachers' assessment of the exam papers.

As the fourth finding of the study, the effect of the number of raters in the group moderation assessment model was examined according to the average Cronbach's alpha reliability values calculated from the total scores given by different numbers of raters to the exam papers. While choosing different numbers of raters, the selection of two, three, four and five raters, respectively, among five raters, was made by a mathematical combination method and the average Cronbach's alpha reliability values were calculated and compared (Graphic 1).



	1st Exam	2nd Exam	3rd Exam
🟦 2 raters	0,923	0,948	0,981
🟥 3 raters	0,944	0,963	0,989
🟩 4 raters	0,956	0,972	0,99
🟪 5 raters	0,969	0,979	0,994

Graphic 1. Comparison of the average Cronbach's alpha values obtained by assessing the exams according to the groups formed by different number of raters.

When we look at the average Cronbach's alpha values obtained from the Cronbach's alpha values of the scores given to the exam papers by the teachers in different number of rater groups formed to examine the students' first, second and third exam papers, we can say that the average Cronbach's alpha values of the scores increase as the number of teachers in the group increases. In addition, when we look at the first, second and third exam assessments of the groups, we can say that the obtained Cronbach's alpha values also increased. The point to note here is the finding that when the number of raters is reduced, the number of evaluation exams must be increased, and if the number of exams is to be reduced, the number of raters must be increased.

4. DISCUSSION, CONCLUSION AND SUGGESTIONS

It is increasingly accepted that the group moderation assessment model can provide teachers with professional learning opportunities in assessment (Adie, 2013; Harlen, 2010; Earle, 2020). Measurement results obtained with open-ended exams contain scoring errors caused by the rater. Sometimes these errors can be seen as scoring out of the answer key even in the case of scoring with the answer key. Therefore, in cases where open-ended exams are to be used especially for important decisions, scoring them independently by two or more raters trained in scoring and determining the average of the raters as the student's score is considered a good solution. In the studies conducted by Lane and Sabers (1989) to determine the scoring reliability, it was determined that as the number of raters increased, the standard deviation decreased, a rater was suitable for group evaluations, and the number of raters should be increased for individual evaluations. In observational assessments, there may be many more than two observers available for any given data set. Because observations are susceptible to human error, a large number of observers is often advantageous for reducing the effects of individual biases in rating. In managerial performance appraisal and clinical diagnosis, for example, the use of three or more raters is not uncommon (London and Wohlers, 1991; Tsui & Ohlott, 1988; Wohlers & London, 1989).

As a result of the study, it was concluded that the items in the mathematics exam measured the mathematics achievement quite consistently, according to the reliability coefficients obtained from the mean scores given by the raters to each item. After the workshops held within the scope of the group moderation assessment model, it was found that there were high and consistent relationships between the exam assessments of the raters. The group moderation assessment model can increase the overall consistency by increasing the inter- rater consistency (McNamara, 1996; Earle, 2020). In addition, it has been stated that there is evidence that it can be effective because it eliminates excessive differences in rater rigidity and increases rater consistency by reducing individual biases (Weigle, 1994). One of the aims of the group moderation assessment model is to increase the consistency between raters and raters by observing factors such as raters' experience, scoring style or scoring preferences (Kim, 2009; Smaill, 2018; Smaill, 2020).

According to the findings, although it was observed that the severity / generosity levels of the five raters in the first test were different from each other, it was concluded that the raters scored in a consistent way with each other in the exams conducted after the group moderation assessment model workshops. This finding shows that the raters were positively influenced by each other and that they formed a reliable evaluation system by making judgments. In addition, it was found that the variability of the raters in the first exam was different from each other, and it was concluded that the raters scored in a consistent way with each other in the exams conducted after the group moderation assessment model workshops. İlhan (2016), in his study, evaluated the students' open-ended exam by four raters and found that there were statistically significant differences between the scores. In order to be able to say that there are no rater errors, there should not be a statistically significant difference between item and test scores obtained from the scoring made by two or more raters (Güler & Teker, 2015).

In the study, according to the applied group moderation assessment model, it was concluded that at least three raters were required to give students' achievement scores more reliably in open-ended exams. Ideally, it was sufficient to have three people in the assessment process, considering that increasing the number of raters will also increase additional time and financial difficulties. Although there have been no previous studies addressing how researchers evaluate multi-rater reliability, several authors have discussed the lack of agreement among researchers on methods for assessing inter-rater reliability among three or more raters (Meister, 1985; Page & Iwata, 1986). Cunningham (1998) stated that if the scoring results are to be used to make very important decisions about students, it is necessary to use two or more independent raters. The scores given by different raters should be averaged and the decisions about the students should be made according to the calculated average. In fact, in cases where there are statistically significant differences between the two raters, a third rater should be consulted (Cunningham, 1998; İlhan & Çetin, 2014). Kamış and Doğan (2017) stated in their studies that it is not possible to select the raters from the universe randomly due to practical conditions and if the rater needs to be increased, this can be achieved by adding a new rater from the institution to the existing raters or by selecting new raters from within the institution.

In conclusion, in the workshops held during the implementation of the group moderation assessment model, it was found that the teachers' knowledge and opinion with each other positively affected the teachers' ability to assess exam papers. After the workshops, although consistency has improved for raters, statistically significant difference s in severity were still found among raters. These results support the notion that the workshops held and the group evaluation model are more successful in helping raters give more predictable scores (i.e. inter-rater reliability) than ensuring that they give the same scores (i.e., inter-rater reliability). Davis (2016) stated that the scoring experience of raters should be increased in order for rater training to be effective. It has been stated that new / novice raters may look more like experienced raters after a few scorings lessons / training (Lim, 2011; Weigle, 1998). Also, Swartz et al. (1999) stated in their study that if the raters are well trained, the level of consistency between raters increases. In addition, there is some evidence in the literature that training for raters is helpful in reducing scoring errors (İlhan & Çetin, 2014). From this point of view, it can be said that the study is similar to this finding. When we consider the workshops and group moderation assessment model in the study as a rater training, it can be said that this method contributed to more reliable scoring of open-ended questions by the raters.

This article increased the relationship between the group moderation assessment model and teacher professional learning in two ways. Firstly, it has shown that involvement in group moderation assessment model can strengthen teachers' assessment capacity. Secondly, group moderation showed that using the principles and practices of the group moderation evaluation model to inform the evaluation processes can enable teachers to create an assessment-focused professional learning community.

In this study, it can be said that increasing the number of raters in written exams increases the reliability. For this reason, more than one rater should be run in the scoring of the exams, the average (or total) of their scores should be taken or the number of exams should be increased to measure student achievement with at least two raters. Because the scores given to open-ended exam results using the answer key are more reliable, exams should be scored using an answer key prepared in advance. In the evaluation of students' studies such as projects or performance assignments, the group moderation assessment model can also be applied to reveal students' achievement in classroom and out-of-class activities more reliably.

References

Adie, L. E. (2013). The development of teacher assessment identity through participation in online moderation. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 91–106.

Aiken, L. R. (2000). *Psychological testing and assessment (10. Edition)*. Boston: Allyn and Bacon.

Allal, L., & Mottier Lopez, L. (2014). *Teachers' professional judgment in the context of collaborative assessment practice.* In C. Wyatt-Smith, V. Klenowski & P. Colbert (Eds.), Designing Assessment for Quality Learning (pp. 151-165). London: Springer (The Enabling Power of Assessment).

Association for Advanced Training, (1988). *Association for Advanced Training in The Behavioral Sciences*. Pub: Los Angeles.

Baykul, Y. (2000). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması [Measurement in Education and Psychology: Classical Test Theory and Practice]*. Ankara, ÖSYM Yayınları.

Benton, T., & Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters: A Cambridge Assessment Publication, 24*, 37–40.

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 92*(1), 81-90.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice 17*(2), 217–34.

Bramley, T., & Vitello, S. (2018). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(1), 43–58.

Büyükkıdık, S., & Anıl, D. (2015). Performansa dayalı durum belirlemede güvenirliğin genellenebilirlik kuramında farklı desenlerde incelenmesi [Examination of reliability in performance-based assessment in different designs in generalizability theory]. *Eğitim ve Bilim, 40*(177), 285-296.

Büyüköztürk, Ş. (2014). *Sosyal Bilimler için Veri Analizi El Kitabı İstatistik, Araştırma Deseni SPSS Uygulamaları ve Yorum [Data Analysis Handbook for Social Sciences Statistics, Research Design SPSS Applications and Interpretation]* (20th ed.). Ankara: Pegem Akademi Yayıncılık.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing**.** *Educational Research and Evaluation*, *13*(5), 101–434.

Clarke, S. (2011). *Formative Assessment in Action Weaving The Elements Together*. Londres: Hodder Murray.

Cunningham, G. K. (1998). *Assessment in the classroom: constructing and interpreting texts*. Psychology Press.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135.

DeLuca, C., & Johnson, S. (2017). Developing assessment capable teachers in this age of accountability. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 121–126.

Doğan C. D., & Anadol, H.Ö. (2017). Genellenebilirlik Kuramında Tümüyle Çaprazlanmış ve Maddelerin Puanlayıcılara Yuvalandığı Desenlerin Karşılaştırılması [Comparison of Patterns in Generalizability Theory with Fully Crossed and Items Nested to Raters]. *Kastamonu Üniversitesi Kastamonu eğitim Dergisi, 25*(1), 361-372.

Downing, S.M. (2009). *Written Tests: Constructed-Response and Selected-Response Formats.* In Downing, S.M. & Yudkowsky, R. (Eds.) Assessment in Health Professions Education (pp. 149-184). New York and London: Routledge.

Earle, S. (2020). Balancing the demands of validity and reliability in practice: Case study of a changing system of primary science summative assessment. *London Review of Education, 18*(2), 221–235.

Evans-Hampton, T. N., Skinner, C. H., Henington, C., Sims, S., & McDaniel, C. E. (2002). An investigation of situational bias: Conspicuous and covert timing during curriculum-based measurement of mathematics across African American and Caucasian students. *School Psychology Review, 31*(4), 529–539.

Gipps, C., & Stobart, G. (2003). *Alternative Assessment* (Vol. 2). Los Angelas, London, New Delhi, Singapore: SAGE Publications.

Gipps, C.V. (1994). *Beyond testing*. London: The Farmer Press.

Gipps, C.V. (1996). *Assessment for learning. In Assessment in transition: Learning, monitoring and selection in international perspective*, ed. A. Little and A. Wolf, 251–61. Oxford: Pergamon.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5*(1), 13-34.

Gravetter, F. J., & Forzano, L. B. (2012). *Research Methods for the Behavioral Sciences (4th ed.)*. Belmont, CA: Wadsworth.

Gronlund, N.E., & Linn, R.L. (1990) *Measurement and Evaluation in Teaching*. McMillan Company, New York.

Güler, N., & Gelbal, S. (2010). A Study Based on Classical Test Theory and Many Facet Rasch Model. *Eurasian Journal of Educational Research, 38,* 108-125.

Güler, N., & Teker Taşdelen, G. (2015). Açık Uçlu Maddelerde Farklı Yaklaşımlarla Elde Edilen Puanlayıcılar Arası Güvenirliğin Değerlendirilmesi [Evaluation of Inter-rater Reliability Obtained by Different Approaches in Open-Ended Items]. *Journal of Measurement and Evaluation in Education and Psychology, 6*(1), 12-24.

Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. *The Curriculum Journal, 16*, 207 - 223.

Harlen, W. (2010). Professional learning to support teacher assessment. In J. Gardner, W. Harlen, L. Hayward, & G. Stobart (Eds.), *Developing teacher assessment* (1st ed). Open University Press.

Humphry, S. M., & Heldsinger, S. (2019). A two-stage method for classroom assessments of essay writing. *Journal of Educational Measurement, 56*(3), 505–520.

Humphry, S. M., & Heldsinger, S. (2020) A Two-Stage Method for Obtaining Reliable Teacher Assessments of Writing. *Frontiers in Education, 5(6),* 1-10.

Hutchinson, C., & Hayward, L. (2005) The journey so far: assessment for learning in Scotland. *Curriculum Journal, 16*(2), 225-248.

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması [Comparison of ability estimations calculated according to classical test theory and multi-faceted Rasch model in measurements made with open-ended questions.]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *31*(2), 346-368.

İlhan, M., & Çetin, B. (2014). Performans Değerlendirmeye Karışan Puanlayıcı Etkilerini Azaltmanın Yollarından Biri Olarak Puanlayıcı Eğitimleri [Rater Trainings as One of the Ways to Reduce Rater Effects Interfering with Performance Evaluation], *Journal of European Education, 4*(2), 29-38.

Kamış, Ö., & Doğan, C. (2017). Genellenebilirlik Kuramında Gerçekleştirilen Karar Çalışmaları Ne Kadar Kararlı? [How Stable Are Decision Studies Performed in Generalizability Theory?]. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 37*(2), 591-610.

Kan, A. (2005). Yazılı yoklamaların puanlanmasında puanlama cetveli ve yanıt anahtarı kullanımının (aynı) puanlayıcı güvenirliğine etkisi. *Eğitim Araştırmaları Dergisi, 5*(20), 166-177.

Kerlinger, F. N. (1992). *Foundations of Behavioral Research*. New York: Harcourt Brace College Publishers.

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

Klenowski, V., & Wyatt-Smith, C. (2013). *Assessment for Education: Standards, Judgement and Moderation.*

Klenowski, V., & Wyatt-Smith, C. (2010). Standards, Teacher Judgement and Moderation in the Contexts of National Curriculum and Assessment Reform. *Assessment Matters, 2*, 107-131.

Lane, S., & Sabers, D. (1989) Use of Generalizability Theory for Estimating the Dependability of a Scoring System for Sample Essays. *Applied Measurement in Education, 2*(3), 195-205.

Lim, G. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*, 543-560.

London, M., & Wohlers, A. J. (1991). Agreement between subordinate and self-ratings in upward feedback. *Personnel Psychology, 44*(2), 375–390.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman

Maxwell, G., & Gipps, C. (1996). Teacher Assessments of Performance Standards: A Cross-National Study of Teacher Judgements of Student Achievement in the Context of National Assessment Schemes. *Application for funding to the ARC: Interdisciplinary and International Research.*

Malone, L., Long, K., & De Lucchi, L. (2004). All things in moderation. *Science and Children, 41*(5), 30-34.

Maxwell, G.S. (2007). *Implications for moderation of proposed changes to senior secondary school syllabuses.* Brisbane: Queensland Studies Authority.

Meister, D. (1985). *Behavioral Analysis and Measurement Methods.* Publisher: Wiley-Interscience.

Moskal, Barbara M., & Leydens, J.A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10), 1-6.

Nalbantoğlu Yılmaz, F., Başusta, B. (2015). Genellenebilirlik Kuramıyla Dikiş Atma ve Alma Becerileri İstasyonu Güvenirliğinin Değerlendirilmesi [Evaluation of Stitching and Removal Skills Station Reliability with Generalizability Theory]. *Journal of Measurement and Evaluation in Education and Psychology, 6*(1), 107-116.

Özçelik, D. A. (1992). *Ölçme ve Değerlendirme [Measurement and Evaluation]*, Ankara, ÖSYM Yayınları. No:2.

Page, T. J., & Iwata, B. A. (1986). *Interobserver agreement: History, theory and current methods.* In A. Poling & R. W. Fuqua (Eds.), Research methods in applied behavior analysis: Issues and advances (pp. 99– 126). New York: Plenum.

Reiner, C. M., Bothell, T. W., Sudweeks, R. R., & Wood, B. (2002). *Preparing effective essay questions: A self-directed workbook for educators*: New Forums Press.

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher, 94*(1), 31-37.

Sadler, D. (1998). Formative Assessment: revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*, 77-84.

Shavelson, R. J., Yin, Y., Furtak, E. M., Ruiz-Primo, M. A., & Ayala, C. C. (2008). *On the role and impact of formative assessment on science inquiry teaching and learning*. In J. Coffey, R. Douglas & C. Stearns (Eds.), Assessing science learning: Perspectives from research and practice. Arlington, VA: NSTA Press.

Shermis, M. D., & Di Vesta, F. J. (2011). *Classroom assessment in action*. Lanham, MD: Rowman & Littlefied.

Smaill, E. (2020). Using involvement in moderation to strengthen teachers' assessment for learning capability. *Assessment in Education: Principles, Policy & Practice,* DOI: 10.1080/0969594X.2020.1777087.

Smaill, E. (2018). *Social moderation: Assessment for teacher professional learning.* Doctoral thesis, University of Otago. https://ourarchive.otago.ac.nz/handle/10523/7850.

Spiller, D. (2012). *Assessment Matters: Self-assessment and peer assessment*. Teaching Development Unit, University of Waikato, New Zealand.

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research and Practice, 15*, 128–134.

Strachan, J. (2002). Assessment in change: Some reflections on the local and international background to the National Certificate of Educational Achievement (NCEA). *New Zealand Annual Review of Education, 11,* 245- 258.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., de Kruif, R. E. L., Reed, M., Brown, T. T., Levine, M. D., & White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Education and Psychological Measurement, 59*, 492–506.

Takunyacı, M. (2016). *Çoktan seçmeli sorulara dayalı olmayan bir kitle matematik sınavı sürecinin değerlendirilmesi: Grup uyumu değerlendirme modeli [Evaluation of a mass mathematics* *exam process not based on multiple choice questions: Group cohesion assessment model]*. Unpublished Doctoral Dissertation, Marmara University, Institute of Educational Sciences.

Tekin, H. (2000). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]* (14th Ed.). Yargı Yayınları, Ankara.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review, 31*(4), 498–513.

Tsui, A. S., & Ohlott, P. (1988). Multiple assessment of managerial effectiveness: Interrater agreement and consensus in effectiveness models. *Personnel Psychology, 41*(4), 779-803. Retrieved from Google Scholar.

Turgut, M.F. (1992). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]* (9th ed.). Ankara: Saydam Matbaacılık.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287.

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2019). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice*, doi: 10.1080/0969594X.2019.1700212.

Wilson, M. (2004). *Assessment, accountability, and the classroom: A community of judgment. In M. Wilson (Ed.), Toward coherence between classroom assessment and accountability.* 103rd yearbook of the National Society for the Study of Education. Chicago, IL: The University of Chicago Press.

Wohlers, A. J., & London, M. (1989). Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self-awareness. *Personnel Psychology, 42*(2), 235–261.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: Examined implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice, 26*(1), 59–74.

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice.* doi:10.1080/ 0969594X.2019.1602027.

Ethics committee approval for this study was obtained from Sakarya University Scientific Research and Publication Ethics Committee, dated 13/01/2021 and protocol number 30.

Statement of Contribution of Researchers to the Article:

1st author contribution rate: 60%

2nd author contribution rate: 40%

Conflict of Interest Statement:

There is no conflict of interest.

Statement of Financial Support or Acknowledgment:

No financial support was received from any institution for this study.