International Journal of Assessment Tools in Education
2024, Vol. 11, No. 3, 507–523
https://doi.org/10.21449/ijate.1438798
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

The effect of rater training on rating behaviors in peer assessment among secondary school students

Nazira Tursynbayeva[1], **Umur Öç**[2], **İsmail Karakaya**[3*]

[1]Khoja Akhmet Yassawi International Kazakh-Turkish University, Faculty of Social and Human Sciences, Department of Pedagogy and Psychology, Turkestan, Kazakhstan
[2]Ministry of National Education, Refahiye District National Education Directorate, Erzincan, Türkiye
[3]Gazi University, Faculty of Gazi Education, Department of Educational Sciences, Ankara, Türkiye

ARTICLE HISTORY
Received: Feb. 17, 2024
Accepted: July. 21, 2024

Keywords:
Peer assessment,
Rating behavior,
Rater training,
Writing skills.

Abstract: This study aimed to measure the effect of rater training given to improve the peer assessment skills of secondary school students on rater behaviors using the many-facet Rasch Measurement model. The research employed a single-group pretest-posttest design. Since all raters scored all students, the analyses were carried out in a fully crossed (s x r x c) pattern. There were three facets in the research: student, rater, and criteria. The study group consisted of 25 seventh-grade students at a public school in Ankara in the 2021-2022 academic year. All 25 students in the study group were instructed to write compositions. The compositions were examined by the researchers, and 10 were selected for peer assessment. Before the experiment, students were asked to evaluate their peers' writing skills according to the rubric developed by the researchers. Then, rater training was given to the students for four weeks. After the rater training, the students were instructed to re-evaluate the writing skills of their peers. In the research, four rater behaviors were examined: rater severity, rater leniency, differentiated rater severity, and differentiated rater leniency. When the research results were examined, it was observed that rater training contributed to reducing severity, leniency, and differentiated severity and leniency behaviors.

1. INTRODUCTION

One of the aims of today's education system is to prepare and support students for daily life. Helping students acquire and develop daily life skills is a major objective of curriculum. One of the practices applied as part of these objectives is the observation and assessment of students across the curriculum. Assessment and evaluation are used to measure the learning outcomes, behavior acquisition, and the effectiveness of teaching programs (Ertürk, 1979).

The proper functioning of the evaluation mechanism allows for quick and effective solutions to potential problems in the system. Monitoring student progress becomes easier, and learning outcomes are more easily and accurately identified. In this way, both the quality of education increases and development is ensured in a way to facilitate and promote adaptation to

*CONTACT: İsmail KARAKAYA ikarakaya@gazi.edu.tr Gazi University, Faculty of Gazi Education, Department of Educational Sciences, Ankara, Türkiye

e-ISSN: 2148-7456

innovations (Çeçen, 2011; Gürlen et al., 2019; İşman & Eskicumalı, 2003; Kurudayioğlu et al., 2008; Turgut & Baykul, 2010; Yaşar, 2017).

Teachers use different evaluation methods when examining the effects of the educational process. If the evaluation methods used are independent of the student, the evaluation process will be incomplete for the student. This is because students usually have more information about their peers' tasks than their teachers (Somervell, 1993). Involving students in the assessment process increases teacher-student and student-student interaction and contributes to the development of students' responsibility-taking behaviors (Keaten & Richardson, 1993). One of the assessment approaches involving active student participation in the assessment process is peer assessment. Peer assessment is the evaluation of classmates according to specified criteria (Boud et al., 1999). Peer assessment allows students to work together effectively (Kutlu et al., 2010).

The biggest problem in educational settings where peer assessment is used is the reliability of the scores obtained (Donnon et al., 2013). When appropriate environments and conditions are not provided, students cannot make objective evaluations and this causes the evaluation to produce incorrect results (Ellington et al., 1997). In addition, the validity of the assessment will be negatively affected as there will be a rater effect on the assessment. Some of the common rater behaviors are rater severity, rater leniency, and bias (Myford & Wolfe, 2003). The tendency of a rater to give lower scores than other raters in the rater group is called rater severity, and the tendency to give higher scores is called rater leniency (Myford & Wolfe, 2004). Rater bias is the tendency of the rater to be sometimes harsh and sometimes generous when scoring students, depending on the characteristics of the students other than the measured characteristic (Knoch et al., 2007). To reduce or eliminate these rater behaviors, it is recommended to use rubrics, use more than one rater, and provide rater training (Andrade, 2005; Hauenstein & McCusker, 2017; Kubiszyn & Borich, 2024; Lumley & McNamara, 1995; Oosterhof, 1999). In the current study, all of the suggested methods were used to make the rating more valid and reliable. Peer assessment involves, by nature, more than one rater. In peer assessment, before creating the relevant assessment tool, the basic behaviors and criteria related to the task are identified with the students, and the expected behaviors of the students are listed. Students should be involved from the first stage of the assessment process. The type of assessment to be used and which learning outcomes will be assessed should be well explained to the students beforehand. The tasks should be appropriate to the level of the students, similar approaches should be used frequently in class, assessment criteria should be prepared together with the students, and possible disagreements should be resolved (Alıcı, 2010; Bushell, 2006; Kutlu et al., 2010; Stiggins & Chappuis, 2005; Woolfolk et al., 2008). After the definitions and explanations about the task are completed, the students should be instructed on how the assessment should be done.

After the assessment tool is created, rater training should be provided to support students in rating objectively. Lack of objectivity in rating is one of the biggest problems encountered during implementation (Donnon et al., 2013). Students' involvement in the rating process supports teaching, influences students' rating behaviors, and contributes to the validity and reliability of the rating. Students' tendency to give a higher score to their close friends or to classmates who are at the top of the class, their failure to fulfill the responsibilities that need to be observed during peer assessment, and their inability to fully comprehend the criteria may negatively affect the peer assessment process (May, 2008). Students' subjective rating behavior may lead to a biased evaluation of the learning process and learning outcomes, and students failing to fulfill their tasks fully may come to the forefront. Studies show that in peer assessment, students may resort to different ways to give each other higher scores and that they may be biased (Greenan et al., 1997; Johnson & Smith, 1997). In addition, all kinds of rater effects can be expected in peer assessment (Farh et al., 1991; Heslin, 2005). Examining the effectiveness of the techniques used to increase objectivity in evaluations using peer assessment

is very important for the reliability of the scores obtained and the validity of the inferences to be made based on the scores. Therefore, providing rater training may contribute to rating validity. When the literature is examined, it is seen that several studies found that rater training contributed significantly to rating accuracy (Bijani, 2018; Congdon & MeQueen, 2000; Fahim & Bijani, 2011; Kondo, 2010; Loignon et al., 2017; Martin & Locke, 2022; May, 2008; Yeşilçınar & Şata, 2021).

Eliminating or reducing undesirable rater behaviors in performance assessment will contribute to the validity, accuracy, and reliability of the results. When the literature is examined, it is seen that there are studies that investigate the effect of rater training on rater behavior in peer assessment among groups at university level and above (Loignon et al., 2017; Martin & Locke, 2022; May, 2008; Yeşilçınar & Şata, 2021). However, there is no study that investigates the effect of rater training on rater behaviors in peer assessment among students at secondary school level. To fill this gap, this study was conducted to determine how rater training given to improve the peer evaluation skills of secondary school students affects their peer rating behaviors.

This research is important to determine the rater behaviors that occur during the process of using peer evaluation and to determine the effect of rater training on eliminating or reducing these behaviors. Focusing especially on the rater behaviors of secondary school students in the peer evaluation process shows the originality of the study. In light of all this information, it was aimed to investigate the effect of rater training with the multi-facet Rasch model in order to provide more objective and accurate scoring in the evaluation of the writing tasks prepared by secondary school students. For this purpose, answers were sought to the following questions.

1) Regarding peer evaluation scores before rater training;
   a) What is the severity and leniency of the raters?
   b) What are the raters' differentiated leniency and severity behaviors?
2) Regarding peer evaluation scores after rater training;
   a) What is the severity and leniency of the raters?
   b) What are the raters' differentiated leniency and severity behaviors?

2. METHOD

2.1. Study's Design

This study employed a single-group pretest-posttest design, aiming to measure secondary school students' rater behaviors when evaluating the writing skills of peers and the effect of rater training on the students' rating behavior in peer assessment, using the many-facet Rasch measurement model. Since each rater scored all students, the analyzes were carried out in a fully crossed pattern. There were three facets in the research: rater, criterion, and student.

2.2. Study Group

The study group consisted of 25 seventh-grade students at a public school in Ankara in the 2021-2022 academic year. The students included in the study were selected according to the following criteria: not having received rater training before, willingness to participate in the study voluntarily, and traceability.

2.3. Data Collection Tools

A writing task and an analytical rubric developed by the researchers were used as data collection tools in the study. During the analytical rubric development process, opinions were taken from three Turkish teachers and two measurement and evaluation experts. The content validity index of the measurement tool was determined using the Lawshe (1975) technique based on expert opinions (CVR=0.99, $p<0.05$). The criteria were arranged to suit the students' levels, and the analytical rubric was finalized. In the rubric, each criterion was evaluated on a four-point scale (1: very unsuccessful; 4: very successful). After the rubric was finalized, validity and reliability studies were conducted. Exploratory factor analysis was conducted to provide evidence for the validity of the rubric. While conducting exploratory factor analysis, the average of the scores

given by the 25 raters to the students' writing tasks was used. Before proceeding with exploratory factor analysis, assumptions such as sample size, multiple normality, linearity and outliers were examined. Çokluk et al. (2021, p. 206) state that when determining the sample size in exploratory factor analysis, the individual/item ratio should be at least 2:1. In the current study, it was determined that the sample size assumption was met because the student-criterion ratio was greater than 2:1 (25:7). Additionally, Guadagnoli and Velicer (1988) criticized the theoretical relationship between sample size and number of items and conducted a Monte Carlo study. They state that even if the number of samples in their study is less than 50, values with a factor loading of 0.80 or more will be sufficient for the sampling assumption. Considering that the factor loadings in the current study are greater than 0.80 (C1= 0.948, C2= 0.945, C3= 0.949, C4= 0.954, C5= 0.922, C6= 0.942, C7= 0.957). It was determined that the number assumption was met. For the multivariate normality assumption, the univariate normality assumption must first be examined (Çokluk et al., 2021, p. 29). After determining that all variables meet the univariate normality assumption (Shapiro-Wilk: p1= 0.77, p2= 0.42, p3= 0.23, p4= 0.10, p5= 0.34, p6= 0.66, p7= 0.10, multiple normality assumption (p1,2,3,4,5,6,7>0.05) was examined. The multiple normality assumption was examined with the help of Scatter Plot Matrix, and it was determined that the multiple normality assumption was met. Providing the multiple normality assumption shows that the relationship between the variables is linear (Büyüköztürk, 2002). Additionally, it was determined that there were no extreme values in the data. After determining that the exploratory factor analysis assumptions were met, the KMO test and Bartlett Sphericity test were performed to determine whether the data were suitable for analysis. The KMO value of the data set was 0.909, and the Bartlett test of sphericity was significant (p<0.00). A KMO test value of 0.90 or above is considered excellent (Hutcheson & Sofroniou, 1999). The fact that the Bartlett Test of Sphericity result is statistically significant is another indication that the data set is suitable for exploratory factor analysis (Field, 2005). This shows that the data set is suitable for exploratory factor analysis. Exploratory factor analysis was conducted by taking the average of the scores given by the raters. As a result of exploratory factor analysis, it was found that the criteria were gathered under one factor, and the explained variance was 89.344%. Factor loadings of each criterion were 0.948, 0.945, 0.949, 0.954, 0.922, 0.942, and 0.957, respectively. Additionally, the Cronbach alpha reliability of the measurements was calculated and found to be 0.98. According to all these results, it can be said that the analytical rubric developed in this study provides valid and reliable results.

2.4. Data Collection Process

The study involved a two-stage data collection process. In the first stage, the analytical rubric to be used in writing skill evaluation was prepared and the rater group was informed about peer assessment, the writing task, and the rubric. In addition, sample applications were shared with the rater group. Ten compositions, selected from those written by the students, were distributed to the students for scoring. Students were given 10 minutes for each composition, 100 minutes in total, for scoring. The students evaluated the compositions written by their peers, and pre-test scores were obtained. In the second stage, the students received rater training two hours a week for a total of four weeks, totaling eight class hours, and then the students were asked to score their peers' compositions once again, and post-test scores were obtained.

2.5. Data Analysis

In the study, ten compositions written by 25 students for a task were selected. These tasks were scored by 25 students according to seven criteria. The average of the scores given by 25 students to each criterion was used in the factor analysis. For the multi-facet Rasch model, the scores given by 25 students to 10 writing tasks were used.

Analyses were performed using the FACET package program. Before proceeding with the analysis, the assumptions of the many-facet Rasch model, including unidimensionality, local

independence, and model-data fit, were examined (Eckes, 2011, p. 124; Farrokhi et al., 2012). As a result of exploratory factor analysis, it was seen that the measurement tool was unidimensional. Meeting the unidimensionality assumption also indicates that the local independence assumption is met (Hambleton et al., 1991). For model-data fit, the ratios of the standardized residuals in the ±2 and ±3 intervals were examined. Linacre (2014) stated that the proportion of standardized residuals outside the ±2 interval should not exceed 5%, and the proportion of standardized residuals outside the ±3 interval should not exceed 1%. In the study, the total number of interactions was 1750 (10 student * 7 criteria * 25 raters), the proportion of standardized residuals outside the ±2 interval was 4.29% (n=75), and the proportion of standardized residuals outside the ±3 interval was 0.74% (n=13). As such, it can be said that model-data fit is achieved, and the inferences to be made in line with the analysis results are valid.

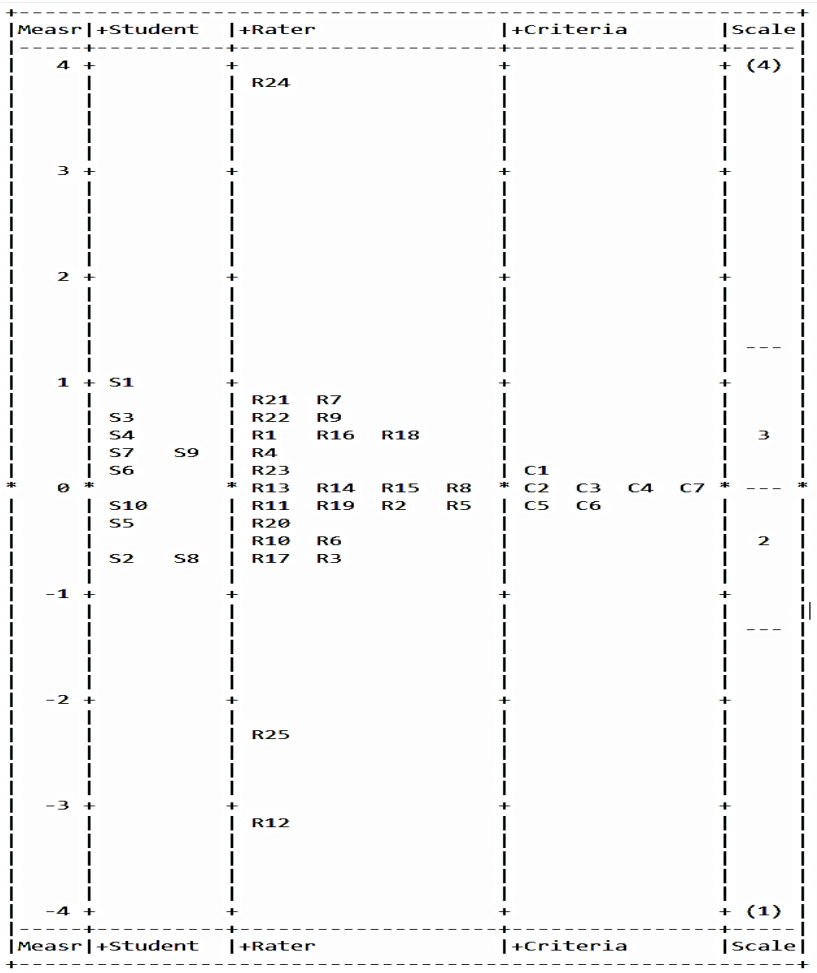
3. RESULTS

The findings obtained in the study are presented under two separate subheadings. The results before rater training (pre-test) are reported under the first, and the results after rater training (post-test) are reported under the second subheading. Under both subheadings, group statistics are given first, followed by individual statistics on a student basis.

3.1. Research Findings Before Rater Training (Pre-Test)

The pre-test calibration map of peer scores, the rater facet measurement report, rater severity and leniency, and biased interactions measured before rater training within the scope of the study are given below.

Figure 1. *Calibration map of peer scores before rater training.*



When the calibration map of peer scores before rater training in Figure 1 is examined, it is seen that the facets are on a logit scale. A high or low logit value has different implications depending on the relevant facet. In the student column, a high logit value at the top of the column indicates a high level of ability, whereas a low logit value at the bottom indicates a low level of ability. In the rater column, the raters with the highest logit values at the top of the column score leniently, while those with the lowest logit value at the bottom score severely. In the criterion column, a high logit value at the top of the column indicates a highly difficult criterion, whereas a low logit value at the bottom indicates low difficulty. To exemplify, when the calibration map is examined, it is seen that the student with the highest ability level in the pre-test is S1, and the students with the lowest ability levels are S2 and S8. The most lenient rater is R24, while the most severe rater is R12. It is also seen that C5 and C6 are the most difficult criteria, while C1 is the easiest. The fact that the student, rater, and criteria facets take values along the negative and positive ends of the logit scale indicates that the students' ability levels, the criteria difficulty levels, and rater rating behaviors are differentiated. The rater facet measurement reports for a detailed examination of rater behaviors are presented in Table 1.

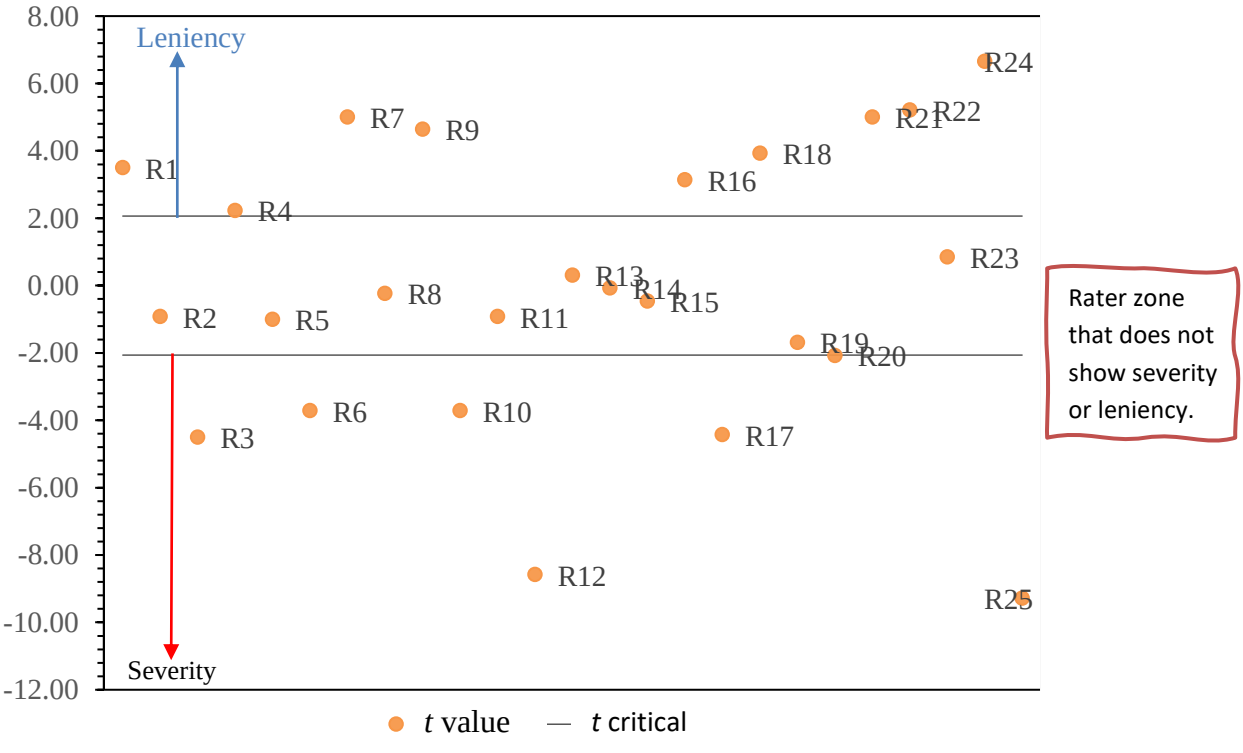
Table 1. *Rater facet pre-test measurements measurement report.*

Rater	Obsvd Average	Fair Average	Logit	Model S.E.	Infit	Outfit
R1	2.96	3.00	0.49	0.14	0.99	1.00
R2	2.47	2.46	-0.12	0.13	0.82	0.79
R3	2.06	2.01	-0.63	0.14	0.98	1.03
R4	2.80	2.83	0.29	0.13	1.39	1.39
R5	2.46	2.45	-0.13	0.13	1.31	1.27
R6	2.14	2.10	-0.52	0.14	0.80	0.76
R7	3.14	3.20	0.75	0.15	0.79	0.80
R8	2.54	2.54	-0.03	0.13	1.30	1.28
R9	3.07	3.13	0.65	0.14	0.95	1.01
R10	2.14	2.10	-0.52	0.14	0.62	0.69
R11	2.47	2.46	-0.12	0.13	1.23	1.21
R12	1.11	1.10	-3.09	0.36	0.94	1.26
R13	2.60	2.61	0.04	0.13	1.13	1.15
R14	2.56	2.56	-0.01	0.13	0.91	0.89
R15	2.51	2.51	-0.06	0.13	0.79	0.77
R16	2.91	2.96	0.44	0.14	0.63	0.66
R17	2.07	2.02	-0.62	0.14	1.10	1.11
R18	3.00	3.05	0.55	0.14	1.11	1.09
R19	2.39	2.37	-0.22	0.13	1.25	1.26
R20	2.34	2.32	-0.27	0.13	0.99	1.02
R21	3.14	3.20	0.75	0.15	0.82	0.87
R22	3.13	3.19	0.73	0.14	0.85	0.94
R23	2.66	2.67	0.11	0.13	0.83	0.84
R24	3.96	3.96	3.86	0.58	0.94	0.71
R25	1.24	1.21	-2.32	0.25	1.01	1.62
Mean	2.56	2.56	0.00	0.17	0.98	1.02
S (population)	0.58	0.61	1.17	0.10	0.20	0.24
S (sample)	0.60	0.62	1.20	0.10	0.21	0.25
Model, Population		RMSE= 0.19 Adj S.D.= 1.16 Separation= 6.00 Strata= 8.33 Reliability= 0.97				
Model, Sample		RMSE= 0.19 Adj S.D.= 1.18 Separation= 6.12 Strata= 8.50 Reliability = 0.96				
Model, Chi-square (fixed effect):		433.1 *df*= 24 *p*= 0.00				
Model, Chi-square (Normal):		21.3 *df*= 23 *p*=0.56				

Table 1 shows the observed and adjusted means, logit values, standard error of logit values, concordance and non-concordance values of the raters before rater training. The logit measures of the raters ranged between -3.09 and 3.86, with a difference of 6.95. A positive value in the logit values of the raters indicates leniency, and a negative value indicates severity behavior. The average infit and outfit values of the raters are close to one. This shows that the model-data fit is good.

It is seen that there are two different models of the rater facet population and sample. If the model includes all possible components of the facet, the "model population" should be interpreted according to the values in the "model sample" row (Linacre, 2014). Accordingly, the values in the "model sample" row were interpreted. It is seen that the discrimination rate (6.12) and reliability index (0.96) are high. The reliability index value calculated for the rater facet shows a reliable difference (Haiyang, 2010). This shows that raters exhibit differentiated severity/leniency behaviors. When Table 1 is examined, it is seen that there are fixed effects and normal Chi-square values for the rater facet. The "normal Chi-square" value should be used to examine whether the facet components represent a randomly selected sample from a normally distributed population, and the "fixed-effect Chi-square" value should be used to examine whether there is a difference between the facet components after allowing for measurement error (Linacre, 2014). Accordingly, the fixed-effect Chi-square value was used to examine whether there was a significant difference in terms of the raters' severity and leniency behaviors. The Chi-square values of the rater facet before rater training were statistically significant $\chi^2(sd) =$ 433.1 (24), p=0.00 < 0.01). This shows that the raters exhibited differentiated behaviors (severity/leniency). After having determined that the raters exhibited differentiated behaviors at the group level, individual student statistics were examined. While examining the raters' behaviors on a student basis, the *t* value was used. After comparing the obtained *t* value with the critical *t* value in the t distribution table, its statistical significance was determined. *t* value was obtained by dividing the difference between the logit value of the rater and the logit mean of all raters by the standard error. The degrees of freedom for the 25 raters before rater training was 24. At a 0.05 level of significance for 24 degrees of freedom, *t* critical was found to be 2.064. The distribution of *t* values for pre-test scores is given in Figure 2.

Figure 2. *Distribution of t values for pre-test scores.*



When Figure 2 is examined, it is seen that 16 (64.00%) of the 25 raters exhibited severity or leniency behavior before rater training. While nine of these raters (36.00%) displayed leniency behavior, seven of them (28.00%) displayed severity behavior. Rater and student interactions were examined to determine differentiated rater severity and leniency at the group level, rater bias in the rater group. Since the Chi-square statistic result of the rater group was significant $\chi^2(sd)$ =535.2 (250), p= 0.00 < 0.01), it was determined that there was a group-level bias effect among the raters. After determining the bias effect at the group level, student-based statistical indicators were examined. In the many-facet Rasch model, a t value outside the ±2 range indicates significance, that is, rater bias (Linacre, 2023, p. 190). Significant interactions for the pre-test are given in Table 2.

Table 2. *Pre-test significant rater-student interactions.*

Rater	Student	Observed Score	Expected Score	Bias (Logit)	Standard Error	t
R1	S8	11.00	16.44	-1.05	0.52	-2.02
R2	S7	12.00	18.42	-1.12	0.48	-2.34
R2	S9	25.00	18.67	1.32	0.59	2.25
R4	S9	12.00	21.11	-1.56	0.48	-3.27
R4	S5	10.00	17.00	-1.45	0.59	-2.44
R4	S2	21.00	14.87	0.99	0.42	2.35
R5	S4	12.00	19.81	-1.34	0.48	-2.81
R5	S5	7.00	14.42	-2.86	1.42	-2.02
R5	S7	24.00	18.31	1.07	0.52	2.08
R6	S7	10.00	15.82	-1.26	0.59	-2.13
R6	S3	24.00	17.98	1.12	0.52	2.18
R6	S4	25.00	17.34	1.53	0.59	2.61
R8	S6	12.00	17.80	-1.02	0.48	-2.14
R8	S5	28.00	15.03	3.69	1.41	2.61
R9	S3	20.00	24.17	-0.85	0.41	-2.09
R10	S8	17.00	11.14	1.10	0.40	2.79
R11	S9	26.00	18.67	1.73	0.71	2.44
R11	S6	26.00	17.25	1.95	0.71	2.75
R12	S5	9.00	7.45	1.51	0.72	2.10
R13	S4	15.00	20.84	-0.94	0.41	-2.29
R13	S5	9.00	15.45	-1.63	0.72	-2.26
R13	S8	25.00	13.75	2.12	0.59	3.62
R14	S7	27000	19.09	2.36	1.00	2.36
R15	S4	28.00	20.23	2.86	1.41	2.03
R17	S10	19.00	12.81	1.03	0.40	2.59
R17	S9	25.00	15.49	1.82	0.59	3.11
R18	S8	23.00	16.80	1.07	0.47	2.27
R19	S10	7.00	14.99	-2.96	1.42	-2.08
R19	S4	26.00	19.27	1.63	0.71	2.30
R19	S6	23.00	16.58	1.10	0.47	2.34
R19	S5	25.00	13.92	2.09	0.59	3.56
R20	S2	17.00	11.96	0.90	0.40	2.29
R21	S3	20.00	24.53	-0.95	0.41	-2.35
R22	S7	18.00	23.14	-0.91	0.39	-2.31
R22	S3	20.00	24.46	-0.93	0.41	-2.29
R23	S9	28.00	20.08	2.88	1.41	2.04
R24	S5	25.00	27.59	-2.00	0.59	-3.42
R25	S2	14.00	7.69	2.46	0.42	5.81

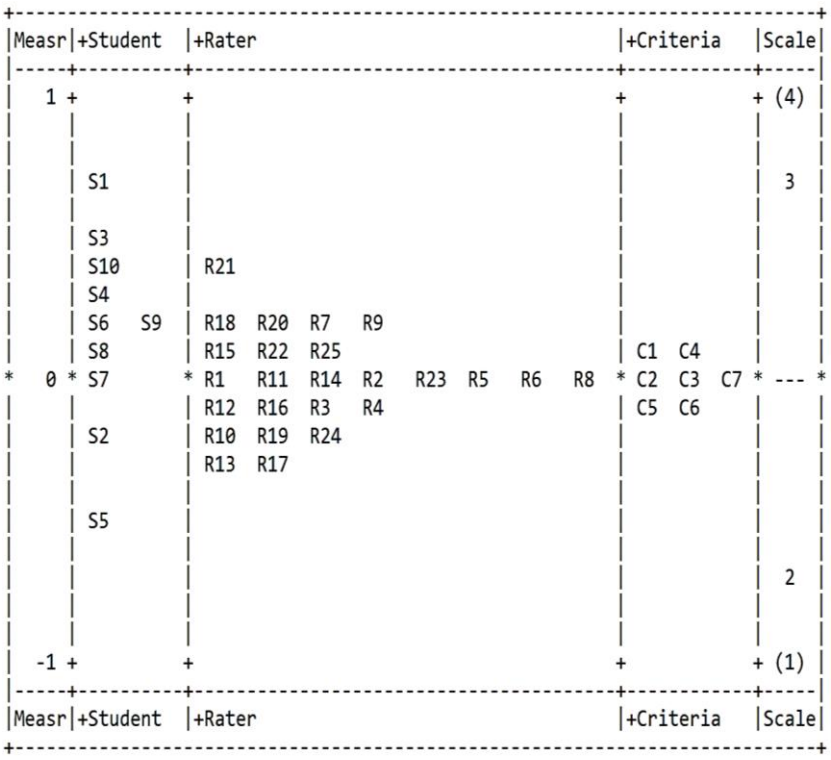
Chi-square = 535.2, *sd*= 250, *p*= 0.00

When Table 2 is analyzed, 38 out of 250 possible interactions between rater and student facets (15.20%) were found to be statistically significant. This indicates that the raters rated some students severely and some students leniently.

3.2. Findings After Rater Training (Post-Test)

The post-test calibration map of peer scores, the rater facet measurement report, rater severity and leniency, and biased interactions measured after rater training within the scope of the study are given below.

Figure 3. *Calibration map of peer scores after rater training.*



When the calibration map of peer scores after rater training in Figure 3 is examined, it is seen that the student with the highest ability level is S1, the student with the lowest ability level is S5, the most lenient rater is R21, the most severe raters are R13 and R17, the easiest criteria are C1 and C4, and the most difficult criteria are C5 and C6.

The measurement reports of the rater facet for a detailed examination of post-test rater behaviors are given in Table 3. Table 3 shows the observed and adjusted means, logit values, standard error of logit values, concordance, and non-concordance values of the raters after rater training. A positive value in the logit values of the raters indicates leniency, while a negative value indicates severity behavior. The average of the infit and outfit values of the raters is 1.00. This shows that the model-data fit is good. The logit measures of the raters vary between -0.29 and 0.36 and the difference is 0.65. The discrimination rate (0.54) and reliability (0.23) are low. The reliability value calculated for the rater facet shows a reliable difference (Haiyang, 2010). This shows that the raters have similar behaviors. After rater training, the fixed-effect Chi-square values of the rater facet were not statistically significant ($\chi^2(sd) = 32.0(24)$, p=0.13>0.01). This is an indication that the raters do not have severity or leniency behavior at the group level. After determining that raters exhibited similar behaviors at the group level, individual statistics on a student basis were examined. The *t* value was used when examining the raters' behaviors on a student basis. After comparing the obtained *t* value with the critical *t* value in the *t* distribution table, its statistical significance was determined. *t* value was obtained by dividing the difference between the logit value of the rater and the logit mean of all raters by the standard

error. The degree of freedom was 24 for the 25 raters after rater training. At a 0.05 level of significance for 24 degrees of freedom, *t* critical was found to be 2.064. The distribution of *t*-values for the post-test scores is given in Figure 4.

Table 3. *Rater facet post-test measurements measurement report.*

Rater	Observed Average	Fair Average	Logit	Model S.E.	Infit	Outfit
R1	2.63	2.63	0.00	0.14	1.07	1.06
R2	2.61	2.62	-0.02	0.14	0.99	0.99
R3	2.54	2.54	-0.11	0.14	1.15	1.16
R4	2.53	2.53	-0.13	0.14	0.88	0.88
R5	2.64	2.64	0.02	0.14	1.35	1.34
R6	2.64	2.64	0.02	0.14	0.82	0.82
R7	2.74	2.75	0.16	0.14	0.51	0.50
R8	2.64	2.64	0.02	0.14	1.36	1.36
R9	2.76	2.76	0.18	0.14	0.79	0.80
R10	2.47	2.47	-0.21	0.14	0.72	0.73
R11	2.60	2.60	-0.04	0.14	1.54	1.55
R12	2.57	2.57	-0.07	0.14	0.92	0.92
R13	2.43	2.43	-0.27	0.14	0.92	0.92
R14	2.61	2.62	-0.02	0.14	1.12	1.11
R15	2.70	2.70	0.10	0.14	0.74	0.74
R16	2.57	2.57	-0.07	0.14	0.99	0.99
R17	2.41	2.41	-0.29	0.14	1.04	1.04
R18	2.76	2.76	0.18	0.14	0.68	0.68
R19	2.47	2.47	-0.21	0.14	1.61	1.58
R20	2.80	2.81	0.24	0.14	1.10	1.11
R21	2.89	2.89	0.36	0.14	0.67	0.68
R22	2.73	2.73	0.14	0.14	0.95	0.95
R23	2.64	2.64	0.02	0.14	1.18	1.19
R24	2.51	2.51	-0.15	0.14	1.22	1.21
R25	2.71	2.72	0.12	0.14	0.61	0.62
Mean	2.63	2.63	0.00	0.14	1.00	1.00
S (population	0.12	0.12	0.16	0.00	0.28	0.27
S (sample)	0.12	0.12	0.16	0.00	0.28	0.28
Model, Population		RMSE= 0.14 Strata= 1.06	Adj. S.D.= 0.08 Reliability= 0.23		Separation= 0.54	
Model, Sample		RMSE= 0.14 Strata = 1.12	Adj. S.D.= 0.08 Reliability = 0.26		Separation= 0.59	
Model, Chi-square (Fixed Effect):		32.0	*df*= 24	*p*= 0.13		
Model, Chi-square (Normal):		13.8	*df*= 23	*p*=0.93		

Figure 4. *Distribution of t values for post-test scores.*

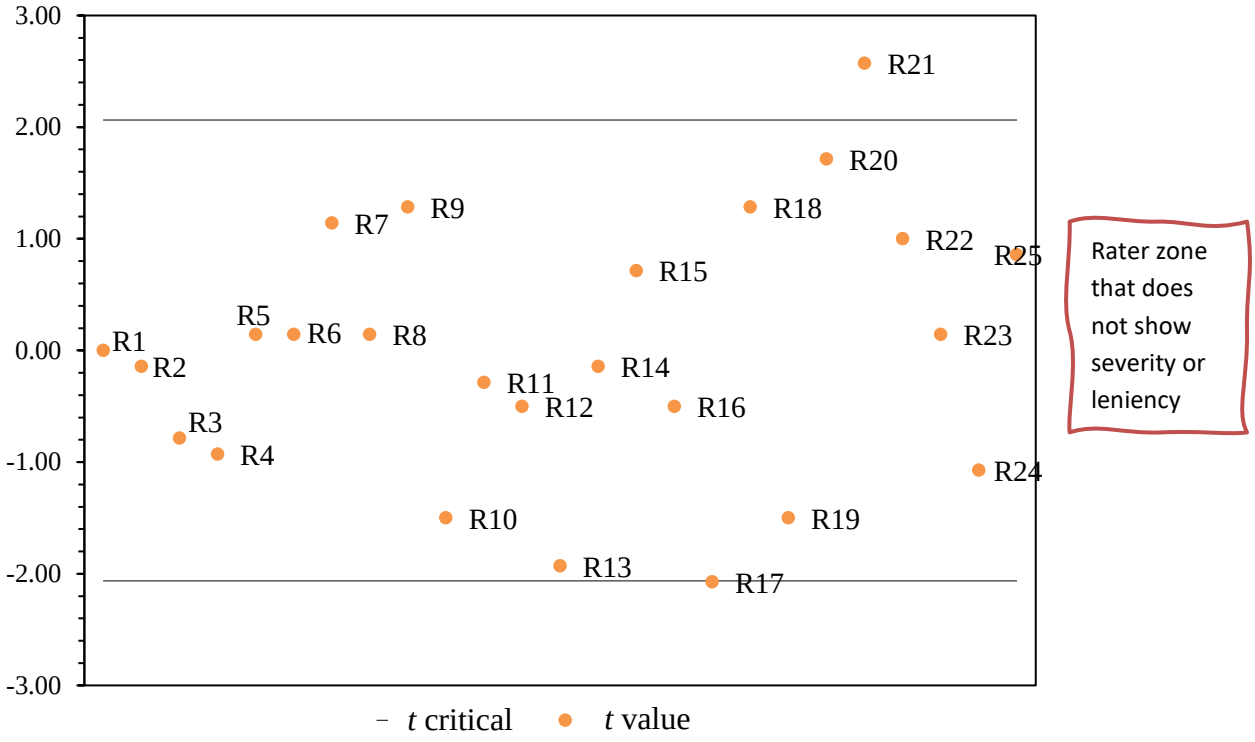


Figure 4 shows that two of the 25 raters (8.00%) exhibited severity or leniency behavior after rater training. One of these raters (4.00%) exhibited leniency behavior, and the other (4.00%) exhibited severity behavior.

The pre-test and post-test statistics of the raters were compared to examine whether there was a statistical difference between rater severity and leniency. For this purpose, *t* statistics, which are indicators of the strictness and generosity of the raters, were compared. Pre-test *t* statistics were compared to post-test *t* statistics as it is an indicator of rater severity/leniency. However, to better observe the impact of rater training, the pre-test t statistical value was differentiated from the raters' post-test *t* statistics. The Mann-Whitney U test results for *t* statistics of pre-test and post-test data are given in Table 4.

Table 4. *Mann Whitney U test results of pre-test and post-test t statistics.*

Test	N	Mean Rank	Sum of Ranks	U	*p*
Pre-test	25	20.06	501.50	176.50	0.008
Post-test	25	30.94	773.50		
Total	50				

When Table 4 is examined, a statistically significant difference is seen in the raters' pre-test and post-test severity and leniency behaviors (U=176.50; p=0.008<0.05), indicating a statistical difference in rater severity/leniency before and after rater training. It can also be said that this difference is in favor of the post-test when considering the decrease in rater effect after rater training.

Rating and student interaction were studied to determine whether the rater had exclusive behavior at the group level. As a result of Chi-square statistics ($\chi^2(sd) = 389.8$ (250), p= 0.00 < 0.01), it was determined that there was a significant group-level bias effect in the rater group. Student-based statistical indicators were studied after the group-level isolation effect was identified. In the many-facet Rasch model, *t* value outside the ±2 range indicates significance, punctuation (Linacre, 2023, p. 190). Post-test significant interactions are shown in Table 5.

Table 5. *Post-test significant rater-student interactions*

Rater	Student	Observed Score	Expected Score	Bias (Logit)	Standard Error	*t*
R1	S10	12.00	19.35	-1.52	0.52	-2.95
R1	S1	26.00	21.03	1.52	0.74	2.07
R3	S1	14.00	20.47	-1.27	0.46	-2.73
R3	S3	14.00	19.44	-1.06	0.46	-2.29
R4	S10	27.00	18.64	2.74	1.02	2.68
R5	S10	25.00	19.45	1.39	0.62	2.25
R8	S3	14.00	20.13	-1.20	0.46	-2.58
R8	S4	14.00	19.37	-1.05	0.46	-2.26
R8	S9	25.00	18.84	1.51	0.62	2.44
R8	S5	21.00	15.19	1.13	0.46	2.47
R10	S1	26.00	19.99	1.74	0.74	2.35
R11	S10	13.00	19.14	-1.23	0.49	-2.54
R11	S9	25.00	18.54	1.57	0.62	2.53
R12	S8	23.00	17.81	1.09	0.51	2.14
R12	S3	27.00	19.63	2.54	1.02	2.49
R13	S8	22.00	16.79	1.04	0.48	2.17
R15	S7	13.00	18.05	-1.03	0.49	-2.11
R15	S4	26.00	19.77	1.78	0.74	2.41
R16	S5	10.00	14.71	-1.26	0.63	-2.01
R16	S4	24.00	18.86	1.17	0.55	2.12
R16	S3	25.00	19.63	1.36	0.62	2.19
R18	S10	15.00	20.24	-1.01	0.45	-2.25
R19	S6	8.00	17.26	-2.98	1.03	-2.90
R19	S7	10.00	16.42	-1.60	0.63	-2.55
R19	S10	28.00	18.23	3.53	1.43	2.47
R19	S4	26.00	18.15	2.09	0.74	2.84
R20	S3	14.00	21.18	-1.42	0.46	-3.05
R20	S2	23.00	17.57	1.13	0.51	2.23
R22	S2	10.00	17.05	-1.72	0.63	-2.74
R22	S10	26.00	20.04	1.73	0.74	2.34
R23	S10	13.00	19.45	-1.29	0.49	-2.66
Chi-square= 389.8 *sd*= 250 *p*= 0.00						

When Table 5 is examined, it is seen that 31 of the possible 250 interactions (12.40%) between the rater and the student facets were statistically significant. This shows that the raters scored some students with severe scores while others with lenient scores.

4. DISCUSSION and CONCLUSION

This study was conducted to determine the effect of rater training, which is one of the methods used to determine and reduce or eliminate rater effect in peer assessment. The many-facet Rasch model was used to determine the rater effect in this study. Pre-test severity and leniency behaviors of the rater group were examined, and as a result, group-level severity and leniency behaviors were observed in the rater group. After the analysis of severity and leniency behaviors at the group level, individual statistics on a student basis were examined. While 16 (64%) of the 25 raters in the rater group were found to be severe or lenient, nine (36.00%) of them were found to have leniency behavior and seven (28.00%) to have severity behavior. Pre-test differentiated rater severity and leniency behaviors at the group level were also included. After the analysis of group-level statistics, the student-level statistics were analyzed. As a result of

the analysis, 38 (15.20%) of the 250 possible interactions between student and rater facets were found to be statistically significant. While 16 of the significant interactions were differentiated rater severity, 22 of them were differentiated rater leniency. These findings are consistent with the studies conducted by Esfandiari and Myford (2013); Farrokhi et al. (2012), Engelhard (1994), Farrokhi and Esfandiari (2011), Karakaya (2015), Şata et al. (2020).

When the post-test severity and leniency of the rater group were examined, it was observed that there was no severity or leniency behavior at the group level. After the analysis of group-level statistics, student-level statistics were analyzed. As a result of the analysis, it was found that two (8.00%) of the 25 raters had severity or leniency behavior: one (4.00%) had rater leniency behavior, and one (4.00%) had rater severity behavior. This may indicate that the two raters may have similar behavior to the pretest.

In addition, a statistically significant difference was found between the raters' pre-test and post-test rater severity and leniency behaviors. This is an indication that rater training was effective in reducing the severity and leniency behaviors of the raters. It was observed that differentiated rater severity and leniency behaviors at the group level continued after rater training. However, only 31 (12.40%) of the 250 possible interactions between the student and rater facets after rater training were found to be statistically significant. While 14 of the significant interactions were differentiated rater severity, 17 were differentiated rater leniency.

Although a decrease in rater effect could be observed after rater training, it did not disappear completely. Many studies investigating the effect of rater training on rater behavior in peer assessment report that rater effect will not change even with feedback or that it will reduce rater behaviors to a certain extent (Berg, 1999; Elder et al., 2005; Knoch, 2011; Knoch et al., 2007; Loignon et al., 2017; Lumley & McNamara, 1995; Lunt et al., 1994; O'Sullivan & Rignall, 2007; Patri, 2002; Wigglesworth, 1993). These studies support the results of this research.

The study sought to explain the possible reasons why rater behaviors did not disappear completely. There are several ways of reducing differential rating inclination and leniency behavior. The first of these methods is to give feedback and rigorous training to the rater. In the study, students did not receive any feedback after rating. Immediate feedback after rating could help raters be more objective when evaluating peers. Knoch (2011) also noted that it would be useful for feedback to raters to be long-term. The lack of feedback in this study may be a cause of bias. There is no standard period in the literature for how long rater training should be given. In this study, students received a total of eight hours of rater training. Giving rater training for an extended period of time may increase the effectiveness of rater training. During rater training, students were given two samples for each criterion. Increasing the number of samples can help students better internalize the criteria. Students (25 students) had limited time to evaluate their peers. This may have caused the raters to misrate some criteria. If the students had had enough time, their scores could have been more objective. Another way could be one-on-one teaching without rater training (Saito, 2008).

The examples used in teaching can make it easier to internalize criteria. The lack of feedback to students and the limited number of samples may have decreased the effect of rater training on rater behavior. Moreover, the task selected for the purpose of this study was persuasive writing. The fact that this type of writing is not included in the Turkish course curriculum may be the reason why rater behaviors have not disappeared.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). **Ethics Committee Number**: Gazi University, 22.04.2022-345966.

Contribution of Authors

Each author has made an equal contribution to the research.

Orcid

Nazira Tursynbayeva https://orcid.org/0000-0002-2165-3276
Umur Öç https://orcid.org/0000-0002-1269-1115
İsmail Karakaya https://orcid.org/0000-0003-4308-6919

REFERENCES

Alıcı, D. (2010). Öğrenci Performansının Değerlendirilmesinde Kullanılan Diğer Ölçme Araç ve Yöntemleri [Other Measurement Tools and Methods Used in the Evaluation of Student Performance (pp. 127-168), Measurement and Evaluation in Education]. *Ankara: Pegem Akademi Yayıncılık*

Andrade, H. G. (2005). Teaching With Rubrics: The Good, the Bad, and the Ugly. *College Teaching*, *53*(1), 27-31. https://doi.org/10.3200/CTCH.53.1.27-31

Berg, E.C. (1999). The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing*, *8*(3), 215-241. https://doi.org/https://doi.org/10.1016/S1060-3743(99)80115-5

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education*, *5*(1), 1460901. https://doi.org/10.1080/2331186X.2018.1460901

Boud, D., Cohen, R., & Sampson, J. (1999). Peer Learning and Assessment. *Assessment & Evaluation in Higher Education*, *24*(4), 413-426. https://doi.org/10.1080/0260293990240405

Bushell, G. (2006). Moderation of peer assessment in group projects. *Assessment & Evaluation in Higher Education*, *31*(1), 91-108. https://doi.org/10.1080/02602930500262395

Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı [Factor Analysis: Basic Concepts and its Use in Scale Development]. *Eğitim Yönetimi: Teori ve Uygulama*, *32*(32), 470-483

Congdon, P.J., & MeQueen, J. (2000). The Stability of Rater Severity in Large-Scale Assessment Programs. *Journal of Educational Measurement*, *37*(2), 163-178. https://doi.org/https://doi.org/10.1111/j.1745-3984.2000.tb01081.x

Çeçen, M. (2011). Türkçe Öğretmenlerinin Seviye Belirleme Sınavı ve Türkçe Sorularına İlişkin Görüşleri [Turkish Language Teachers' Views About Level Determination Exam and Turkish Lesson Questions]. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi 8(15), 201-212.* https://dergipark.org.tr/en/pub/mkusbed/issue/19555/208689

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2021). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları* [*Multivariate Statistical SPSS and LISREL Applications for Social Sciences*] (6 ed.). Pegem Akademi Yayıncılık https://doi.org/10.14527/9786055885670

Donnon, T., McIlwrick, J., & Woloschuk, W. (2013). Investigating the Reliability and Validity of Self and Peer Assessment to Measure Medical Students¡¯ Professional Competencies. *Creative Education*, *4(6)*, Article 32932. https://doi.org/10.4236/ce.2013.46A005

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual Feedback to Enhance Rater Training: Does It Work? *Language Assessment Quarterly*, *2*(3), 175-196. https://doi.org/10.1207/s15434311laq0203_1

Ellington, H., Earl, S., & Cowan, J. (1997). Making effective use of peer and self assessment. *Innovations in Education and Training International*, *32*, 175-178.

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. https://doi.org/https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Ertürk, S. (1979). Program development in education (3rd Edition). Yelkentepe Publications.

Esfandiari, R., & Myford, C.M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing*, *18*(2), 111-131. https://doi.org/https://doi.org/10.1016/j.asw.2012.12.002

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *International Journal of Language Testing*, *1*(1), 1-16.

Farh, J.-L., Cannella, A.A., & Bedeian, A.G. (1991). The Impact of Purpose on Rating Quality and User Acceptance. *Group & Organization Studies*, *16*(4), 367-386. https://doi.org/10.1177/105960119101600403

Farrokhi, F., & Esfandiari, R. (2011). A Many-facet Rasch Model to Detect Halo Effect in Three Types of Raters. *Theory & Practice in Language Studies*, *1*(11).

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, *34*(1), 79-101.

Field, A. (2005). Reliability analysis. *Discovering Statistics Using spss. 2nd Edition, Sage, London*.

Greenan, K., Humphreys, P., & McIlveen, H. (1997). Developing transferable personal skills: part of the graduate toolkit. *Education + Training*, *39*(2), 71-78. https://doi.org/10.1108/00400919710164161

Guadagnoli, E., & Velicer, W.F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265-275. https://doi.org/10.1037/0033-2909.103.2.265

Gürlen, E., Boztunç Öztürk, N., & Eminoğlu, E. (2019). Investigation of the Reliability of Teacher, Self and Peer Evaluations at Primary School Level Using Generalizability Theory. Journal of Measurement and Evaluation in Education and Psychology, 10(4), 406-421.

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics*, *33*(2), 87-102.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hauenstein, N.M.A., & McCusker, M.E. (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment*, *25*(3), 253-266. https://doi.org/https://doi.org/10.1111/ijsa.12177

Heslin, P.A. (2005). Conceptualizing and evaluating career success. *Journal of Organizational Behavior*, *26*(2), 113-136. https://doi.org/https://doi.org/10.1002/job.270

Hutcheson, G.D., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage.

İşman, A., & Eskicumalı, A. (2003). Eğitimde Planlama ve Değerlendirme [Planning and Evaluation in Education] (4th Edition). *Istanbul: Değişim Yayınları*

Johnson, C., & Smith, F. (1997). Assessment of a complex peer evaluation instrument for team learning and group processes. *ACCOUNTING EDUCATION-GREENWICH-*, *2*, 21-40.

Karakaya, İ. (2015). Comparison of Self Peer and Instructor Assessments in the Portfolio Assessment by Using Many Facet Rasch Model. *Journal of Education and Human Development*, *4*(2).

Keaten, J.A., & Richardson, M.E. (1993). A Field Investigation of Peer Assessment as Part of the Student Group Grading Process.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior - a longitudinal study. *Language Testing*, *28*(2), 179-200. https://doi.org/10.1177/0265532210384252

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*(1), 26-43. https://doi.org/https://doi.org/10.1016/j.asw.2007.04.001

Kondo, Y. (2010). Examination of Rater Training Effect and Rater Eligibility in L2 Performance Assessment. *Journal of Pan-Pacific Association of Applied Linguistics*, *14*(2), 1-23.

Kubiszyn, T., & Borich, G.D. (2024). *Educational testing and measurement*. John Wiley & Sons.

Kurudayioğlu, M., Şahin, Ç., & Çelik, G. (2008). Türkiye'de Uygulanan Türk Edebiyatı Programı'ndaki Ölçme ve Değerlendirme Boyutu Uygulamasının Değerlendirilmesi: Bir Durum Çalışması [Evaluation of the Application of Measurement and Evaluation Dimension in Turkish Literature Program Implemented in Turkey: A Case Study]. *Ahi Evran University Kırşehir Eğitim Fakültesi Dergisi, 9*(2), 91-101. https://dergipark.org.tr/en/pub/kefad/issue/59525/856034

Kutlu, Ö., Doğan, C.D., & Karakaya, İ. (2010). Öğrenci başarısının belirlenmesi performansa ve portfolyoya dayalı durum belirleme [Determining student achievement based on performance and portfolio assessment].Ankara: Pegem Akademi Yayıncılık

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563-575. https://doi.org/https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J.M. (2014). *A user's guide to FACETS: Rasch-model computer programs* (Vol. 18). http://www.winsteps.com/manuals.htm

Linacre, J.M. (2023). *Facets computer program for many-facet Rasch measurement*. Winsteps.com.

Loignon, A.C., Woehr, D.J., Thomas, J.S., Loughry, M.L., Ohland, M.W., & Ferguson, D.M. (2017). Facilitating peer evaluation in team contexts: The impact of frame-of-reference rater training. *Academy of Management Learning & Education*, *16*(4), 562-578. https://doi.org/10.5465/amle.2016.0163

Lumley, T., & McNamara, T.F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, *12*(1), 54-71. https://doi.org/10.1177/026553229501200104

Lunt, H., Morton, J., & Wigglesworth, G. (1994). Rater behaviour in performance testing: Evaluating the effect of bias feedback. 19th annual congress of Applied Linguistics Association of Australia: University of Melbourne. July,

Martin, C.C., & Locke, K.D. (2022). What Do Peer Evaluations Represent? A Study of Rater Consensus and Target Personality [Brief Research Report]. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.746457

May, G.L. (2008). The Effect of Rater Training on Reducing Social Style Bias in Peer Evaluation. *Business Communication Quarterly*, *71*(3), 297-313. https://doi.org/10.1177/1080569908321431

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., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of applied measurement*, *5*(2), 189-227.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. *IELTS Collected Papers: Research in speaking and writing assessment*, 446-478.

Oosterhof, A. (1999). *Developing and using classroom assessments*. ERIC.

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, *19*(2), 109-131. https://doi.org/10.1191/0265532202lt224oa

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, *25*(4), 553-581. https://doi.org/10.1177/0265532208094276

Somervell, H. (1993). Issues in Assessment, Enterprise and Higher Education: the case for self-peer and collaborative assessment. *Assessment & Evaluation in Higher Education*, *18*(3), 221-233. https://doi.org/10.1080/0260293930180306

Stiggins, R., & Chappuis, J. (2005). Using Student-Involved Classroom Assessment to Close Achievement Gaps. *Theory Into Practice*, *44*(1), 11-18. https://doi.org/10.1207/s15430421tip4401_3

Şata, M., Karakaya, İ., & Erman Aslanoğlu, A. (2020). Evaluation of University Students' Rating Behaviors in Self and Peer Rating Process via Many Facet Rasch Model [Üniversite Öğrencilerinin Öz ve Akran Puanlama Sürecinde Puanlama Davranışlarının Many Facet Rasch Modeli ile İncelenmesi]. *Eurasian Journal of Educational Research*, *20*(89), 25-46. https://dergipark.org.tr/en/pub/ejer/issue/57497/815802

Turgut, M.F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]* (Vol. 2). Ankara: Pegem Akademi Yayıncılık

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305-319. https://doi.org/10.1177/026553229301000306

Woolfolk, A.E., Hoy, A.W., Hughes, M., & Walkup, V. (2008). *Psychology in education*. Pearson Education.

Yaşar, M. (2017). Ölçme ve değerlendirmenin önemi [The importance of measurement and evaluation]. *Pegem Citation Index*, 2-8.

Yeşilçınar, S., & Şata, M. (2021). Examining Rater Biases of Peer Assessors in Different Assessment Environments. *International Journal of Psychology and Educational Studies*, *8*(4), 136-151. https://dergipark.org.tr/en/pub/pes/issue/65718/1020683