Eurasian Journal of Educational Research 89 (2020) 25-46

Eurasian Journal of Educational Research
www.ejer.com.tr

Evaluation of University Students' Rating Behaviors in Self and Peer Rating Process via Many Facet Rasch Model

Aslihan ERMAN ASLANOGLU[1], Ismail KARAKAYA[2], Mehmet SATA[3]

ARTICLE INFO

Article History:
Received: 20 Jan. 2020
Received in revised form: 29 Mar. 2020
Accepted: 19 May 2020
DOI: 10.14689/ejer.2020.89.2

Keywords
Peer assessment, self-assessment, rater bias, alternative assessment, Many-Facet Rasch Model

ABSTRACT

Purpose: When self and peer assessment methods become commonly used in the teaching process, the most important problem turns out to be the reliability of the ratings acquired from these sources. Increasing the rater reliability has great importance in the performance evaluation for the reliability of the measurement. This study aimed to determine rater behaviors university students display in the process of self and peer assessment. The research was based on a descriptive model. The participants were 58 students at the Guidance and Psychological Counseling Program in 2017-2018 academic year at a foundation university in Ankara.

Findings: Many Facet Rasch Model (MFRM) analysis was applied, and no statistically significant difference of raters' severity and leniency behaviors in the ratings was observed in terms of gender, but there was a statistically significant difference based on the rater types (self and peer). The raters seemed to be more lenient in self-assessments. The study also showed that while raters showed central tendency behavior on individual level, they did not show such tendency at the group level. It was concluded that individuals' ratings are more biased than group ratings when they evaluate group performance.

Implications for Research and Practice: Some of the raters had differentiating rating behaviors based on the groups. The teacher candidates made systematic mistakes in the performance evaluation process and showed behaviors that had negative effect on the validity of the rating. It is important for the raters to conduct studies to reduce the scoring bias of the raters.

[1] Corresponding Author, Faculty of Education, Ufuk University, TURKEY, e-mail: aslihanerman@yahoo.com
ORCID: 0000-0002-1364-7386
[2] Faculty of Education, Gazi University, TURKEY, e-mail: ikarakaya2002@gmail.com
ORCID: https://orcid.org/0000-0003-4308-6919
[3] Faculty of Education, Agri Ibrahim Cecen University, TURKEY, e-mail: msata@agri.edu.tr
ORCID: https://orcid.org/0000-0003-2683-4997

Introduction

It is clear that the main aim of higher education has inclined to support students to turn them into critical thinkers on their own professional practices, problem solvers and reflective practitioners (Falchikov & Goldfinch, 2000; Kwan & Leung, 1996). Individuals' gaining and developing those skills has also been the focus of programs of instruction. Thereby, observation and evaluation of those aforementioned skills by programs of instruction is in question. Classical assessment tools implemented for this purpose remain incapable in measurement of those mentioned features. This new understanding sees the participation of students in the evaluation of learning process also as important. Hence, this situation has highlighted the use of new evaluation approaches (Bushell, 2006; Falchikov & Goldfinch, 2000). Unlike traditional approaches, students are not only passive information recipients in new assessment approaches. Students' gaining higher level cognitive skills such as critical and creative thinking and problem solving constitutes the basis of this approach (Kutlu, Yildirim & Bilican, 2009). In performance evaluation which gained importance with new approaches, instead of choosing any of the options offered; the student should generate the answer herself/himself (Unal & Ergin, 2006). Thus, unlike multiple-choice tests that relate the student to retrieve information from memory, performance evaluation is based on the process of structuring knowledge actively (Moore, 2009). In this process, students should have an opportunity to interact with their peers and teachers. Thus, it becomes possible for students to structure the information and share the structured information. Assessment and evaluation are instruments for learning that is becoming increasingly desirable to ensure students to take responsibility for their own learning by involving them in this process (Dochy & McDowell, 1997).

Self-assessment and peer assessment are considered as important evaluation approaches for students to take responsibilities for their learning, and it is suggested to encourage students to participate actively in teaching process using these assessments.

Self- assessment is defined as a formative assessment process in which students evaluate their own studies in accordance with predetermined criteria and goals, and increase the quality of the studies by making arrangements according to the results of these evaluations (Andrade, Du & Mycek 2010). With the help of self-assessment, students take more responsibility for their own learning and actively participate in the process of "assessment for learning" (Ballantyne, Hughes & Mylonas, 2002; Matsuno, 2006). Self-assessment is determined by the teachers and minimizes the problems that may arise from the assessment based on the criteria that the students are not generally informed so that they allow the students to evaluate their own studies and learn new things from their mistakes. Puhl (1997) interpreted its biggest contribution as "one of the important skills that should be developed for students to take with them when they leave school and then use them for lifelong learning".

In peer assessment, which is another method of assessment, students are active participants in the whole process as in self-assessment. Peer assessment is defined as an arrangement for students of similar status to consider and take into account the

value or quality of the products of each other's learning output (Topping, Smith, Swanson & Elliot, 2000), and in this respect, it is seen as a planning job (Topping, 1998). In line with this planning, peer assessment serves to "both formative assessment which is based on observation with the aim of giving feedback and summative assessment which is based on placement in terms of determining success" (Temizkan, 2009). Studies show that students find peer assessment more useful in their learning (Landry, Shoshanah & Newton, 2015). Peer evaluation may also be one of the guiding elements in group work, which is necessary for today's business life. Accordingly, the peer assessment practice carried out in group work may contribute to the success of individuals as it may increase the responsibility of individuals.

Self-assessment and peer assessment make the assessment procedure more systematic and formal. Students compare their learning to their peers' and make inferences about their own learning. Also, as the number of evaluators increases, it is possible to get to know the student in a multi-faceted way. In other words, students will have a multidimensional feedback on the quality of their work more than to the extent that they can be evaluated by one instructor with classical methods (Millar, 2003).

When self and peer assessment methods are used in the teaching process, the most important problem is the reliability of the scores obtained (Donnon, Mcllwrick & Wololoschuk, 2013). Increasing the interrater reliability is of great importance in the performance evaluation to increase the reliability of the measurement. The results obtained from the performance measurement can be valid only if the scores are reliable (Jonsson & Svingby, 2007). Therefore, in the performance evaluations, it is necessary to examine the interrater consistency before evaluating the results (Cakici-Eser & Gelbal, 2013). The factors affecting the performance of the student are called rater effects (Farrokhi, Esfandiari & Vaez Dalili, 2011). In the process of self and peer assessment, various rater effects can be observed due to the raters.

Rater effects interfering with performance evaluation and affecting the reliability are examined under different titles such as rater severity and leniency, central tendency behavior, halo effect, range restriction (Saal, Downey & Lahey, 1980), bias and inconsistency (Myford & Wolfe, 2004). Research shows that peer scoring is made more severe but in self-assessment, raters are more lenient in scoring (Falchikov & Boud, 1989; Farrokhi, Esfandiari & Dalili 2011; Farrokhi, Esfandiari & Schaefer 2012; Karakaya, 2015; Lejk & Wyvill 2001; Topping, 2003). Nonetheless, the literature suggested various methods to be utilized such as scoring rubric to reduce the errors originating from raters (Author & Co-author, 2003; Andrade 2005; Oosterhof, 2003), training of raters (Hauenstein & McCusker, 2017; Lumley & McNamara, 1995; Rose, 2006), inclusion of more than one rater to the process (Kubiszyn & Borich, 2013), and including such practices more in classroom (Author, 2017; Bushell, 2006; Topping, 2003; Zhang, 2008), thus there would be less concern about the reliability of scores. In this study, both more than one rater and scoring rubric have been employed for more reliable measurement in the process of self and peer assessment of the students' performances.

The researchers recommend the Many-facet Rasch Model (MFRM) to determine the reliability of peer and self-assessment scores and eliminate the limitations of classical approaches (Baird, Hayes, Johnson, Johnson & Lamprianou, 2013; Kim, Park & Kang, 2012; Linacre, 1996; Lunz, Wright & Linacre, 1990). In assessing the performance of the students by MFRM, the factors that may affect the students' scores are not limited to the skill levels of individuals or the difficulty levels of the items used in the measurement process. Factors related to raters can also lead to variability in student performance scores (Johnson & Lamprianou, 2013). This feature of MFRM makes it a viable option for performance assessments affected by rater behavior (Mulqueen, Baker & Dismukes, 2000). MFRM is also considered to be a more powerful psychometric model according to the classical test theory in terms of features such as being able to identify the interactions between different sources of error (Haiyang, 2010), taking into account more than one source of error at the same time, producing higher ability estimates for validity (Ilhan, 2016), providing information at the individual level rather than at the group level for raters or individuals whose performances are being evaluated (Barkaoui, 2008).

When the studies about MFRM are examined, it is observed that some of the researchers (Guler, 2008; Macmillan, 2000; Sudweeks, Reeve & Bradshaw, 2005) benefited from MFRM in comparative studies with other theories. Some of these studies aim to determine the success of individuals and the severity/leniency of the raters (Akin & Basturk, 2012; Basturk, 2008; Engelhard & Stone, 1998; McNamara & Adams, 1991; Weigle, 1998; Weigle, 1999), some of them aim to investigate rater bias and factors affecting it (Aryadaust, 2015; Cetin & Ilhan 2017, Farrokhi & Esfandiari, 2011; Saito, 2008; Schaefer, 2008; Wolfe, 2004), and some others aim to investigate rater sources-self, peer and teacher-(Farrokhi, Esfandiari, & Dalili, 2011). This research considered the participation of teacher candidates in the assessment process (self and peer assessment) as contributing to improve their scoring behaviors and make the teaching processes more efficient. In addition, the research aimed to contribute to the literature concerning teacher candidates' scoring behaviors during the assessment of individual performance.

We emphasize that it is significant to use self and peer assessment in performance evaluation. It is also important to determine the errors committed by scorers during the assessment of individual performance when self and peer assessments are concerned. Therefore, the present study pointed to the type of evaluation for the errors and uncovered the scoring behaviors involved in the assessment. Besides, the use of Rasch Model, which provides a deeper and broader framework in performance evaluation, promoted the robustness of the study.

This study aimed to determine which rater behaviors university students were manifested during self and peer rating process with the help of MFRM. For this purpose, the questions sought to be answered in the study were as follows;

1. Do the severity and leniency behaviors of the raters differ significantly according to their gender?

2. Do the severity and leniency behaviors of the raters differ significantly according to the rater type (self and peer)?
3. Do the central tendency behaviors (rating categories, criteria, and groups) of the raters differ significantly from each other?
4. Do the raters show biased rating behavior?

Method

Research Design

The study showed a descriptive type of quantitative research feature as it aimed to reveal the rating behaviors of the prospective teachers in the process of scoring the research proposals they prepared. Since all the raters evaluated all group work, a fully crossed design was used. Due to the description of an existing situation in the research, there were five surfaces including the raters, gender of the raters, group work, rater type (self and peer), and criteria. The study aimed to examine the rating behaviors of self and peer assessment during the performance evaluation process. Both group-level statistics and individual-level statistics were conducted to determine rater attitudes the raters displayed.

Participants

The participants of the study were 58 volunteers among the students who took Scientific Research Methods class in 2017-2018 academic year at the Guidance and Psychological Counseling Program of the Faculty of Education at a foundation university in Ankara. Due to the fact that the participants were teacher candidates who were enrolled in the course taught by one of the researchers of this study at the time of the data collection, no permission was obtained, and participation in the study was on a voluntary basis.

Research Instruments and Procedures

The data included in the study were collected by an analytical scoring rubric (ASK) developed by the researchers. ASK was developed to evaluate any scientific research proposal. Firstly, expert opinions were taken for the measurement tool developed as a draft. The measurement tool took its final form in accordance with opinions and suggestions. Accordingly, the criteria of the measurement tool were determined as the statement of the problem, method, findings and result/comment. Each criterion of ASK was rated using a quadruple rating (rather inadequate, 0; quite adequate, 3).

After the application of the ASK, studies were conducted to determine the validity and reliability of the measurements. Exploratory factor analysis (EFA) was used for evidence of the validity of the measurements. The case of whether the assumptions of the exploratory factor analysis were met were examined, which demonstrated that the necessary assumptions were met. The KMO value for the corresponding data set was 0.775, Bartlett's test of sphericity was significant, all criteria of the scoring rubric were normally distributed, and there was no outlier or missing value. The mean score of 58 students in 12 group studies was calculated while AFA was performed. The results of

the EFA showed that the criteria in the ASK were collected under a single factor, and the explained variance was 93.121%. The factor loadings of the criteria for the relevant data set were as follows; 0.946; 0.973; 0.982; 0.958.

The reliability coefficient (ω) proposed by McDonald (1999) was used for the reliability of the measurements. Since the factor loadings of the variables were different from each other in the present study, it was preferred to use McDonald's coefficient for the more consistent predictions of such measurements (Osburn, 2000). As a result of the analysis, McDonald's coefficient was found to be 0.982 (%95 Confidence Interval: 0.952-0.994). According to this result, it can be argued that the measurements obtained from the ASK developed to measure student group work provided valid and reliable results.

Data Analysis

In the analysis of the data, MFRM was used. Analyses were conducted using FACETS software. The analysis had some assumptions. Compensating these assumptions served the validity of inferences based on the analysis results. Unidimensionality was examined as the first assumption, and it showed that the measurement tool had a single dimension as a data collection tool. Ensuring unidimensionality was considered as an indication that local independence was also met, and no action was taken for local independence. Finally, model data compliance was investigated. For model data compliance, the number of standardized residuals outside the ± 2 range should not be more than 5% of the total number of observations, and the standardized residual values outside the ± 3 range should not be more than 1% of the total number of data (Linacre, 2017). It was observed that the model data compliance was provided for the current study as the total number of observations was 2784 (58 x 12 x 4), the standardized residual values outside the ± 2 range were 116 (4.17%), and the standardized residual values outside the ± 3 range were 28 (1.01%) in this study.

Results

Within the scope of this study, rater severity, rater leniency, central tendency and rater bias behaviors were examined.

Rater Severity and Leniency

Before evaluating the self and peer assessments of the raters, the infit and outfit values of each rater were examined. It was determined that 4 out of 62 of the raters had poor compliance values (outliers) and were excluded from the analysis. Upon the exclusion, the analysis was repeated. The analytic outcomes of the gender of the raters in the evaluation of the group work (measurement report) are presented in Table 1.

Table 1

MFRM Analysis Outcome (Measurement Report) Regarding the Gender of Raters

Gender	Observed Average	Fair-M Average	Measure	Model S.E.	Infit	Outfit
Female	2.35	2.63	0.07	0.04	1.01	1.10
Male	2.36	2.59	-0.07	0.08	0.92	0.96
Mean	2,35	2.61	0.00	0.06	0.96	1.03
S (Population)	0.00	0.02	0.07	0.02	0.04	0.07
S (Sample)	0.00	0.03	0.10	0.03	0.06	0.10
Model, Population: RMSE =0.06 Adj (True) S.D. =.03 Separation = 0.53 Strata= 1.04 Reliability = 0.22						
Model, Sample: RMSE =0.06 Adj (True) S.D. =0.08 Separation =1.25 Strata = 2.00 Reliability = 0.61						
Model, Fixed (all same) chi-square: 2.60 d.f. = 1 significance (probability) = 0.11						

P.S. S.D: standard deviation, d.f.: degree of freedom, RMSE: root mean square error

Table 1 shows that the calculated separation rate, strata and reliability for the sample were low. These low values were considered to be an indicator of similar rater behaviors of male and female raters, in other words, their behavior of similar ratings/evaluations in the process of evaluation of individual performance. When the fixed chi-square value of the male and female raters to determine whether the ratings of male and female raters differed was evaluated, it was found as not statistically significant ($\chi 2(df) = 2.60(1)$, significance = 0.11>0.01). According to this result, in the process of determining the status of the group work, the rater severity and leniency showed no statistically significant difference between male and female raters.

After determining that the gender of the raters was not statistically significant in the performance evaluation process, the significance of the rater type (self and peer) in the performance evaluation process was examined.

Table 2

MFRM Analysis Outcome (Measurement Report) Regarding Rater Type

Rater Type	Observed Average	Fair-M Average	Measure	Model S.E.	Infit	Outfit
Self	2.72	2.80	0.90	0.16	1.37	2.30
Peer	2.32	2.35	-0.90	0.04	0.97	0.96
Mean	2.52	2.57	0.00	0.10	1.17	1.63
S (Population)	0.20	0.23	0.90	0.06	0.20	1.67
S (Sample)	0.28	0.32	1.27	0.09	0.29	1.95
Model, Population : RMSE = 0.12 Adj (True) S.D. = 0.89 Separation = 7.59 Strata = 10.45 Reliability = 0.98						
Model, Sample: RMSE = 0.12 Adj (True) S.D. = 1.26 Separation = 10.78 Strata = 14.71 Reliability = 0.99						
Model, Fixed (all same) chi-square: 117.20 d.f. = 1 significance (probability) = 0.00						

P.S. S.D: standard deviation, d.f.: degree of freedom, RMSE: root mean square error

Table 2 demonstrates the group level statistics, which indicated that the calculated separation rate, strata and reliability for the sample were high. It means that the levels of severity and leniency of the self and peer ratings were different in the process of evaluating the group work. The fixed chi square which was applied to determine whether the severity and leniency levels of self and peer ratings differ statistically showed that it was significant ($\chi 2(df) = 117.20(1)$, significance = 0.00<0.01). When the self and peer logit values (level of severity and leniency) were examined, it was observed that the raters had more lenient behavior in self-assessment while they showed more severe behavior during the process of peer assessment. Moreover, the standard errors of the self-assessments of the raters were higher than the peer ratings, so the reliability of the self-assessments was lower. The examination of concordance values showed that the outfit values of the self-assessment ratings were not within the acceptable limits, in other words, the rating given by the raters was outlier.

After examining the severity and leniency behaviors of self and peer ratings, severity and leniency behavior of each rater was examined. The output of the MFRM analysis for the rater facets was given in Table 3.

Table 3

MFRM Analysis Outcome (Measurement Report) Regarding Rater Facets

Rater No	Logit Value	Standard Error	Infit	Outfit	Observed Agreement	Expected Agreement	t-score
057	4.69	0.34	0.91	0.75	55.10	51.20	4.412
056	4.46	0.32	0.91	0.93	50.20	51.30	3.969
051	4.03	0.30	1.43	1.34	47.80	51.80	2.800
038	4.03	0.31	0.86	0.67	60.20	53.50	2.710
013	3.90	0.30	0.86	1.10	53.50	52.50	2.367
015	3.90	0.30	0.55	0.59	59.20	52.50	2.367
021	3.81	0.30	0.85	0.67	55.50	52.50	2.067
062	2.61	0.26	0.80	0.77	52.60	49.10	-2.231
019	2.60	0.26	1.34	1.22	41.40	48.90	-2.269
055	2.60	0.26	1.20	1.26	43.90	49.20	-2.269
054	2.62	0.25	0.76	0.76	48.50	46.90	-2.280
045	2.43	0.25	0.91	0.95	46.50	47.50	-3.040
058	2.29	0.25	0.92	0.90	47.70	46.60	-3.600
050	1.37	0.24	0.66	0.67	27.50	33.40	-7.583
Mean	3.19	0.27	0.99	1.07			
S(Population)	0.53	0.02	0.22	0.45			
S(Sample)	0.54	0.02	0.22	0.46			
Model, Population: RMSE = 0.27 Adj (True) S.D. = 0.46 Separation = 1.67 Strata = 2.56 Reliability = 0.74							
Model, Sample: RMSE = 0.27 Adj (True) S.D. = 0.46 Separation = 1.69 Strata = 2.58 Reliability = 0.74							
Model, Fixed (all same) chi-square : 215.80 df = 57 significance (probability) = 0.00							
Model, Random (Normal) chi-square: 45.50 df = 43 significance (probability) = 0.84							
Expected interrater agreement percentage = %50.90 Absolute agreement percentage =%51.70							

P.S.: Only raters whose t-scores are significant were included.

Table 3 presents the high calculated separation rate, strata and reliability for the sample. This means that the severity and leniency behaviors of the raters were

different in the performance evaluation process. Among 58 students/raters who evaluated the group work, 14 raters (7 severe, 7 lenient) showed severity and leniency behaviors (see t-scores in Table 3). The performance evaluation process showed that the fixed chi-square test applied for the statistical significance of the raters' severity and leniency behaviors was meaningful ($\chi 2(sd) = 215.80(57)$, significance = 0.00<0.01).

Central Tendency Behavior

Central tendency behavior is frequently encountered in the performance evaluation process. For the third question of the study, raters' central tendency behaviors were examined. First, the group level statistics and then individual level statistics were examined. One of the group-level statistics was category statistics. Table 4 presents the rating category (rating scale) statistics in this study.

Table 4
Category Statistics Regarding the Measurement Tool Used in the Evaluation of Group Work

Rating Categories	Frequency	%	Cumulative %	Average logit measure	Expected logit measure	Outfit
0	15	%1	%1	0.13	-0.04	1.0
1	258	%9	%10	0.79	0.83	1.0
2	1243	%45	%54	2.01	2.00	1.2
3	1268	%46	%100	3.33	3.33	1.0

Analyzing the rating category statistics in Table 4, it was seen that the raters preferred categories of 3 and 4 more, and barely used categories of 0 and 1. Two possible reasons for this may be the result of centralized behavior or individual performance (of the group work) being at the medium-level. The category statistic, which was one of the group-level statistical indicators for determining the real cause of this situation, was not sufficient by itself. Therefore, other statistical indicators at the group level such as the group-level statistics in the measurement reports of the group and criteria surfaces should also be examined. First, the measurement report regarding the surface of the criterion is given in Table 5.

Table 5

MFRM Analysis Output Regarding Criteria Surface (Measurement Report)

Criteria	Observed Average	Fair-M Average	Logit Value	Standard Error	Infit	Outfit
Criteria 1	2.44	2.69	0.32	0.07	1.00	1.14
Criteria 2	2.43	2.68	0.27	0.07	1.03	1.18
Criteria 3	2.35	2.61	0.00	0.07	1.03	1.05
Criteria 4	2.18	2.44	-0.58	0.07	0.91	0.92
Mean	2.35	2.61	0.00	0.07	0.99	1.07
S (Population)	0.10	0.10	0.36	0.00	0.05	0.10
S (Sample)	0.12	0.11	0.41	0.00	0.06	0.11

Table 5 Continue

Criteria	Observed Average	Fair-M Average	Logit Value	Standard Error	Infit	Outfit
Model, Population: RMSE =0.07Adj (True) S.D.=0.35 Separation =4.88 Strata = 6.84 Reliability = 0.96						
Model, Sample: RMSE = 0.07 Adj (True) S.D. = 0.41 Separation = 5.66 Strata = 7.88 Reliability = 0.97						
Model, Fixed (all same) chi-square: 10.90 d.f = 3 significance (probability) = 0.00						
Model, Random (normal) chi-square : 2.90 d.f = 2 significance (probability) = 0.23						

P.S. S.D: standard deviation, d.f.: degree of freedom, RMSE: root mean square error

Table 5 shows that the compliance values for the criteria were within the acceptable range and the standard error values were low. This indicates that all criteria did not impair the model - data compliance. In addition, high values of the separation rate, strata and reliability indicate that the criteria can successfully distinguish the performance of group work. In the performance evaluation process, the fixed chi-square test was meaningful in that the criteria statistically distinguish the group work from each other ($\chi 2(df) = 104.90(3)$, significance = 0.00<0.01). That is, the raters did not show central tendency behavior in the performance evaluation process concerning group work. In addition, when the category possibilities related to criteria were examined, it was observed that the categories of the criteria successfully distinguished group performances from each other. The possibilities for the categories of the criteria are given in Figure 1.

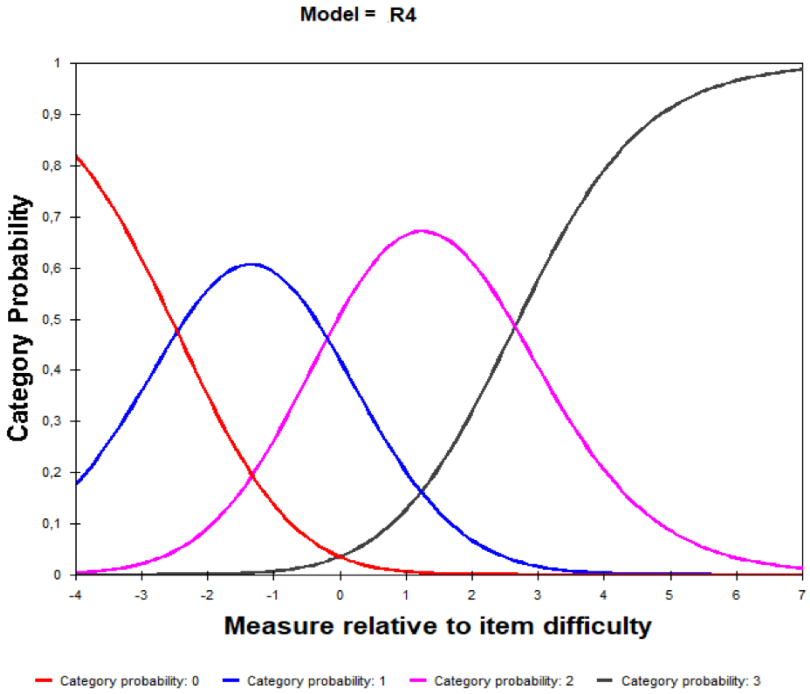

Figure 1. *Category Possibilities*

After examining the criterion surface, it was determined that there was no central tendency behavior at the group level, group surface, which gives information about the group work. MFRM analysis output regarding group surface is presented in Table 6.

Table 6

MFRM Analysis Output (Measurement Report) Regarding Group Surface

Criteria	Observed Average	Fair-M Average	Logit Value	Standard Error	Infit	Outfit
Group2	2.80	2.91	1.80	0.17	1.00	1.86
Group1	2.71	2.86	1.27	0.15	1.02	1.13
Group5	2.65	2.82	1.00	0.14	0.82	0.76
Group3	2.60	2.78	0.78	0.13	0.78	0.76
Group4	2.60	2.78	0.76	0.13	0.96	1.03
Group7	2.49	2.70	0.35	0.12	0.99	1.12
Group6	2.39	2.62	0.03	0.12	1.31	1.28
Group8	2.22	2.47	-0.50	0.12	1.11	1.09
Group9	2.18	2.43	-0.64	0.11	0.84	0.84
Group10	2.05	2.29	-1.06	0.11	0.83	0.85
Group12	1.82	2.06	-1.75	0.11	1.01	1.00
Group11	1.71	1.97	-2.03	0.11	1.13	1.14
Mean	2.35	2.56	0.00	0.13	0.98	1.07
S (Population)	0.34	0.30	1.16	0.02	0.15	0.29
S (Sample)	0.36	0.32	1.21	0.02	0.15	0.30

Model, Population: RMSE = 0.13 Adj (True) S.D. = 1.15 Separation = 9.00 Strata = 12.34 Reliability = 0.99
Model, Sample: RMSE = 0.13 Adj (True) S.D. = 1.21 Separation = 9.41 Strata = 12.88 Reliability = 0.99
Model, Fixed (all same) chi-square: 1015.20 d.f = 11 significance (probability) = 0.00
Model, Random (normal) chi-square: 10.90 d.f = 10 significance (probability) = 0.37

P.S. S.D.: standard deviation, d.f.: degree of freedom, RMSE: root mean square error

Table 6 displays that the separation rate, strata and reliability regarding the group surface were high. In other words, the group performances were distinguished successfully as for that their ability levels. The fixed chi-square test applied for successful distinguishing of group work according to their performances was significant ($\chi 2(df) = 1015.20(11)$, significance = 0.00<0.01). The results indicated that the students/raters did not show group level central tendency behavior in the process of evaluating the group work. The lack of the central tendency behavior at the group level does not guarantee that it will not occur at the individual level. The examination of infit and outfit values of each of the first statistical raters at the individual level pointed out that all the raters had the compliance values within the acceptable range. Secondly, the calculated category statistics for each rater were examined. It was found out that 18 out of 58 raters showed the central tendency behaviors at the individual level, and 12 of those raters performed this behavior on category 2 and 6 of them on category 1. In other words, it was determined that the majority of the raters who showed central tendency behavior preferred a score above the average.

Rater Bias Behavior (Differentiating Rater Severity and Rater Leniency)

The fourth research question determined whether or not the raters showed rating bias behavior. Rater bias emerge in two different ways; differentiating severity and leniency. It is defined as a behavior that occurs frequently in the performance evaluation process and decreases validity directly. It is important to determine rater bias in the performance evaluation studies. One of the major advantages of the MFRM analysis in practice is that it provides evidence for rater bias by using the interaction effects between the surfaces included in the model. Since there were five surfaces in this study, a total of 10 interactions occurred on these surfaces. However, only rater behaviors were taken into consideration, so rater x group work interactions were included. When bias analysis was applied in the MFRM analysis, the t-value, the degree of freedom, the bias size, and the significance values were calculated for the related interactions. Firstly, group-level statistics were analyzed. The analyses demonstrated that the chi-square test performed to determine whether the rater bias occurred at the group level was significant ($\chi 2(df)$ =1048.50(696), significance = 0.00<0.01). According to this result, rater behaviors appeared at the group level during the performance evaluation process.

After it was determined that rater bias occurred at group level, individual level statistics were examined. A t-value was calculated for each element of the rater x group interactions. As a general rule, it is accepted that the interaction element which has outside ± 2 range t-value shows the rater bias (Linacre, 2017, s.218). Since 58 raters made status identification of 12 group work, a total of 696 (58x12) interactions occurred in the current study. As a result, 69 out of 696 possible interactions (%9.91) were statistically significant. Of the 69 individual significant interactions which emerged during the evaluation of group work, rater severity, and leniency behaviors which differentiate based on the sign of t-values were determined. 14 of the 69 significant interactions in the present study were differentiating rater severity while 55 of them were differentiating rater leniency.

Discussion, Conclusion and Recommendations

Rating results of individuals indicate that severity and leniency behaviors show a significant difference according to self and peer ratings. According to self and peer ratings performed in the process of determining the performances, 14 out of 58 students showed severity or leniency behaviors (7 severe, 7 lenient). It can be stated that approximately 25% of the students exhibited this behavior. Compared to a similar study by Farrokhi et al. (2012), the results of this study showed relatively less severity and leniency behaviors. The results of this study are similar to the findings of Engelhard (1994), Farrokhi and Esfandiari (2011), and Karakaya (2015) regarding the severity and leniency behaviors obtained by self and peer rating types. Based on the value obtained from this study, one needs to consider some points from Myford and Wolfe (2003) that proposed to decrease the severity and leniency behaviors of teacher candidates. The fact that there is no significant difference on severity and leniency behaviors regarding the gender in the rating process shows that students exhibit similar levels of behavior.

Central tendency behavior can be described as different raters' usage of rating categories divergently. In other words, some of the raters overuse extreme categories while some of them overuse medium categories (Engelhard, 1994). Regarding the third sub-problem of this study, the rating categories of the raters' central tendency behaviors were examined according to the criteria and groups. According to the findings, some raters showed the central tendency behaviors on individual basis whereas the same phenomenon was not observed in the group. Hence, it shows that group performances were distinguished successfully according to their skills level.

It was also observed that 18 out of 58 raters showed central tendency behaviors on individual level. 12 of those raters performed this behavior on rating category 2 and 6 of them on rating category 1. It indicated that individuals preferred a score that is above average. This can be interpreted as these individuals' using the rating categories in a different way. In other words, these raters used extreme categories more excessively than the other raters. Other raters may tend to overuse the medium categories (Engelhard, 1994).

In the rating process, rater bias can provide important information about the validity. Whether or not the raters made a valid rating were examined by observing the rater bias. Rater bias was first examined by taking a look at individual x group interaction. It was found that there was a different rating, which means biased rating, in the group level. This leads us to the conclusion that individuals made biased rating when evaluating group performance as for groups. For the ratings at the individual level, rating bias occurs in only 69 out of 696 (9.91 %) possible interactions. This makes it necessary for individuals to use scoring rubrics more carefully and be a part of rating education for upcoming ratings. The rater training is important in terms of eliminating the extreme differences in the rater severity and increasing the internal consistency of the rater by reducing the individual prejudices of the raters. (Weigle, 1994).

The use of peer and self-assessments in higher education enables effective learning to take place by making students participate actively in the course and to take responsibility of their learning. In addition, educators can have the opportunity to make multiple evaluation of the students, because as the number of evaluators increases, it will be possible to get more images about the student and recognize them in a multi-faceted way. In other words, students will have a multidimensional feedback on the quality of their work more than to the extent that they can be evaluated by one instructor with classical methods. Despite these benefits, peer and self-assessment have some limitations. Early in the list of these limitations, there is reliability of the ratings. Taking this effect from the raters on individual performance into consideration contributes to the validity and reliability of the measurements and evaluations. In this regard, the present study aimed to contribute to the validity and reliability of the evaluations of the students' performance by examining the effects of the rating in the process of evaluating the assignments, which are the products of the group work of the students in the higher education.

We acknowledged some limitations in this study. First of all, research showed more than 30 rater behaviors in the process of performance evaluation; however, the

present study took into account the most common rater behaviors. The second limitation is that the raters in this study were people who have not had a prior scoring experience. Lastly, since this research was carried out focusing on 'the skills of preparation of a research report', the results would not be generalized to the universe. The study revealed that there was no significant difference of raters' severity and leniency behaviors in the ratings based on gender. The fact that 14 of the individuals exhibited severity and leniency behaviors showed that these raters were composed of both men and women. For this reason, it may be suggested that both genders are to be included in the rating training process regarding severity and leniency. Studies that investigate the effect of gender on performance evaluation report that gender has no significant effect (Porter & Shen, 1991; Winke, Gass & Myford, 2012). Van-Trieste's (1990) study reported that male scorers graded female performance higher while female scorers graded male performance higher. One of the reasons that we observed no statistically significant difference between male and female raters in the study was that the measured performance belonged to the groups rather than individuals, and the groups were composed of men and women.

In addition, in the rating process, the raters had differentiated severity and leniency behaviors based on the self and peer rater types. This shows that individuals behave differently when evaluating their performance or their peer's performance. This is the reason that the studies for self and peer assessment should receive more attention for raters to act more objectively. Especially, studies can be carried out within a program for self-scoring.

In the research, central tendency behavior at individual level was observed, though there was none at group level. This can translate as the individuals' preference of extreme and medium rating categories more. Therefore, it can be suggested that the studies towards the raters' more careful usage of scoring rubric should be dwelled on. In the context of the last sub-problem in the study, it was concluded that some of the raters had a differentiating rating behavior based on the groups. In this respect, it was observed that teacher candidates made systematic mistakes in the performance evaluation process and showed behaviors that had a negative effect on the validity of the rating. In other words, the rating bias of the raters decreases the validity of the rating. For this reason, it is important for the raters to conduct studies to reduce the scoring bias of the raters. Training on performance evaluation can contribute to the decrease of the rater bias of pre-service teachers and improve the validity and reliability of the assessment. In addition, we believe that it is important to provide in-service teachers with training and seminars to decrease rater bias with their scoring behavior.

References

Akin, O. & Basturk, R. (2012). Keman egitiminde temel becerilerin Rasch olcme modeli ile degerlendirilmesi [The evaluation of the basic skills in violin training by many facet Rasch model]. *Pamukkale University Journal of Education,* 31(1), 175-187. Retrieved from https://dergipark.org.tr/pauefd/issue/11112/132860

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

Andrade, H., Du, Y. & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assesment in Education Principles Policy and Practice,* 17(2), 199-214. https://doi.org/10.1080/09695941003696172

Baird, J. A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability. A Comparative exploration from the perspectives of generalisability theory, Rash model and multilevel modelling*. Oxford: University of Oxford for Educational Assessment. Retrieved from http://dera.ioe.ac.uk/17683/1/2013-01-21-marker-effects-and-examination-reliability.pdf

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment and Evaluation in Higher Education*, 27, 427-441. https://doi.org/10.1080/0260293022000009302

Basturk, S. (2008). Ogretmenlik uygulamasi dersinin uygulama ogretmenlerinin görüslerine dayali olarak degerlendirilmesi [Evaluation of teaching practicum course based on the mentors' opinions]. *Educational Sciences and Practice,* 7(14), 93-110. Retrieved from http://ebuline.com/turkce/arsiv/14_7.aspx

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

Cakici-Eser, D. & Gelbal, S. (2013). Genellenebilirlik kurami ve lojistik regresyona dayali hesaplanan puanlayicilar arasi tutarligin karsilastirilmasi [Comparison of interrater agreementcalculated with generalizability theory and logistic regression]. *Kastamonu Education Journal,* 21(2), 423-438. Retrieved from http://www.kefdergi.com/pdf/21_2/21_2_2.pdf

Cetin, B., & Ilhan, M. (2017). An Analysis of Rater Severity and Leniency in Open-ended Mathematic Questions Rated Through Standard Rubrics and Rubrics Based on the SOLO Taxonomy. *Education and Science*, 42(189), 217-247. https://doi.org/10.15390/EB.2017.5082

Dochy, F., & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation*, 23, 279-298. https://doi.org/10.1016/S0191-491X(97)86211-6

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(6A), 23-28. https://doi.org/10.4236/ce.2013.46A005

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/10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G., & Stone, G.E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement*, 58(2), 179-196. https://doi.org/10.1177/0013164498058002003

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70 (3), 287-322. https://doi.org/10.3102/00346543070003287

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, 59, 395-430. https://doi.org/10.3102/00346543059004395

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), 1531-1540. https://doi.org/10.4304/tpls.1.11.1531-1540

Farrokhi, F., Esfandiari, R. & Dalili, M.V. (2011). Applying the Many-Facet Rasch Model to detect centrality in self-Assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal* 15 (Innovation and Pedagogy for Lifelong Learning), 70-77. Retrieved from https://pdfs.semanticscholar.org/dd21/ba5683dde8b616374876b0c53da376c10ca9.pdf

Farrokhi, F., Esfandiari, R. & Schaefer, E. (2012). A Many-Facet Rasch Measurement of differential rater severity/leniency in self assessment, peer assessment, and teacher assessment. *Journal of Basic and Applied Scientific Research*, 2 (9), 8786-8798. Retrieved from https://jalt-publications.org/files/pdf-article/jj2012a-art4.pdf

Guler, N. (2008). *Klasik test kurami, genellenebilirlik kurami ve Rasch modeli uzerine bir arastirma [A research on classical test theory generalizaibility theory and rasch model]*. Unpublished thesis, Hacettepe Universitesi, Ankara.

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, 253-266. https://doi.org/10.1111/ijsa.12177

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, 2(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Karakaya, I. (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), 182-192. https://doi.org/10.15640/jehd.v4n2a22

Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly*, 29(4), 346-365. https://doi.org/10.1123/apaq.29.4.346

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement: Classroom application and practice*. Hoboken, NJ: John Wiley & Sons, Inc.

Kutlu, O., Yildirim, O. & Bilican, S. (2009). Ogretmenlerin dereceli puanlama anahtarlarina iliskin tutum olcegi gelistirme calismasi [Study of attitudes scale development aimed at scoring rubrics for primary school teachers]. *Journal of Yuzuncu Yil University Faculty of Education*, 6(2), 76-88. Retrieved from https://dergipark.org.tr/yyuefd/issue/13712/166014

Kwan, K., & Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment and Evaluation in Higher Education*, 21, 205-215. https://doi.org/10.1080/0260293960210301

Landry, A., Shoshanah, J. & Newton, G. (2015). Effective use of peer assessment in a graduate level writing assignment: A case study. *International Journal of Higher Education*, 4(1)., 38-41. https://doi.org/10.5430/ijhe.v4n1p38

Lejk, M. & Wyvill, M. (2001). The Effect of the inclusion of self-assessment with peer-assessment of contributions to a group project. *Assessment and Evaluation in Higher Education*, 26(6), 551-61. https://doi.org/10.1080/02602930120093887

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. *Objective measurement: Theory into practice*, 3, 85-98. Retrieved from https://files.eric.ed.gov/fulltext/ED364573.pdf

Linacre, J.M. (2017). *A user's guide to FACETS: Rasch-model computer programs*. Chicago: MESA Press.

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

Lunz, M. E., Wright, B. D. & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

McDonald, M. B. (1999). Seed deterioration: Physiology, repair and assessment. *Seed Science and Technology*, 27 (1), 177-237. Retrieved from https://ci.nii.ac.jp/naid/10025267238/

McNamara, T. F., & Adams, R. J. (1991). Exploring rater behavior with Rasch techniques. *Language Testing Research Colloquium,* 1-29. Retrieved from https://files.eric.ed.gov/fulltext/ED345498.pdf

Millar, J. (2003). Gender, poverty and social exclusion. *Social Policy and Society*, 2(3), 181 - 188. https://doi.org/10.1017/S1474746403001246

Moore, B.B. (2009). Consideration of rater effects and rater design via signal detection theory. Unpublished Doctoral Dissertation. Columbia University, New York.

Mulqueen, C., Baker, D., & Dismukes, R.K., (2000). Using multi facet Rasch analysis to examine the effectiveness of rater training. *15th Annual Conference for the Society for Industrial and Organizational Psychology,* https://doi.org/10.1037/e540522012-001

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. Retrieved from http://jampress.org/

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

Oosterhof, A. (2003). *Developing and using classroom assessment*. USA: Merrill/Prentice Hall.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological methods*, 5(3), 343. http://dx.doi.org/10.1037/1082-989X.5.3.343

Porter, D., & Shen, S. (1991). Sex, status and style in the interview. *The Dolphin*, 21(2), 117-128. https://doi.org/10.1002/pssa.2211280113

Puhl, C. A. (1997). Develop, not judge: Continuous assessment in the ESL classroom. *Forum Magazine*, 35(2), 2-9. Retrieved from https://eric.ed.gov/?id=EJ593288

Saal, F. E. , Downey, R. G. & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413-428. https://doi.org/10.1037/0033-2909.88.2.413

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493. https://doi.org/10.1177/0265532208094273

Sudweeks, R. R., Reeve, S.& Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9(3), 239-261. https://doi.org/10.1016/j.asw.2004.11.001

Temizkan, M. (2009). Akran degerlendirmenin konusma becerisinin gelistirilmesi uzerindeki etkisi[The effect of peer assessment on the development of speaking skill]. *Mustafa Kemal University Journal of Social Sciences Institute,* 6(12), 90-98. Retrieved from http://sbed.mku.edu.tr/article/view/1038000386

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3), 249-276. https://doi.org/10.3102/00346543068003249

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility, optimising new modes of assessment: *In Search of Qualities and Standards Innovation and Change in Professional Education*, 1, 55-87. https://doi.org/10.1007/0-306-48125-1_4

Topping, K. J., Smith, E. F., Swanson, I.& Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education*, 25(2), 149-169. https://doi.org/10.1080/713611428

Unal, G., & Ergin, O. (2006). Bulus yoluyla fen ogretiminin ogrencilerin akademik basarilarina, ogrenme yaklasimlarina ve tutumlarina etkisi [Academic of students in science teaching through invention effect on successes, learning approaches and attitudes]. *Journal of Turkish Science Education,* 3(1), 36-52. Retrieved from http://www.tused.org/internet/tufed/arsiv/v3/i1/metin/tufedv3i1s3.pdf

Van-Trieste, R. F. (1990). The relation between Puerto Rican university students' attitudes toward Americans and the students' achievement in English as a second language. *Homines*, 13–14, 94–112.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-87. https://doi.org/10.1177/026553229801500205

Weigle, S. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6 (2), 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11 (2), 197-223. https://doi.org/10.1177/026553229401100206

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231-252. https://doi.org/10.1177/0265532212456968

Üniversite Öğrencilerinin Öz ve Akran Puanlama Sürecinde Puanlama Davranışlarının Many Facet Rasch Modeli ile İncelenmesi

Atıf:

Aslanoglu-Erman, A., Karakaya, I., & Sata, M. (2020). Evaluation of university students' rating behaviors in self and peer rating process via many facet Rasch model. *Eurasian Journal of Educational Research 89, 25-46,* DOI: 10.14689/ejer.2020.89.2

Özet

Problem Durumu: Yükseköğretimin temel amacının, öğrencileri, kendi mesleki uygulamaları üzerinde eleştirel düşünen, problem çözen, yansıtıcı uygulayıcılar haline getirmelerine destek vermeye yöneldiği açıktır (Falchikov & Goldfinch, 2000; Kwan & Leung, 1996). Bireylerin bu becerileri kazanması ve geliştirmesi öğretim programlarının da odak noktası haline gelmiştir. Dolayısıyla öğretim programlarının belirtilen bu becerileri izlemesi ve değerlendirmesi söz konusudur. Bu amaç için uygulanan klasik ölçme araçları sözü edilen özelliklerin ölçülmesinde yetersiz kalmaktadır. Bu yeni anlayış öğrenme sürecinin de değerlendirilmeye öğrencilerin katılmasını önemli görmektedir. Bu durum ise yeni değerlendirme yaklaşımlarının kullanılmasını ön plana çıkarmıştır (Bushell, 2006; Dochy, 2001; Falchikov ve Goldfinch, 2000). Öğrencilerin öğrenmelerinde, sorumluluklarını almaları için öz değerlendirme ve akran değerlendirme önemli değerlendirme yaklaşımları olarak görülmekte ve bu değerlendirmelerin kullanılarak öğrencilerin öğretime aktif olarak katılmalarının teşvik edilmesi önerilmektedir. Öğretimde öz ve akran değerlendirmelerinin kullanılması önemi yadsınamayacak bir yarar sağlamaktadır. Çünkü değerlendiricilerin sayısı arttıkça, öğrenciye ilişkin daha fazla resim elde ederek onu çok yönlü tanımak mümkün olabilecektir. Başka bir deyişle öğrenciler, tek bir öğretim elemanının klasik değerlendirme yöntemlerinden daha fazla değerlendirebileceği ölçüde, yaptıkları çalışmaların kalitesi hakkında çok yönlü bir geribildirime sahip olurlar (Millar, 2003). Öğretim sürecinde öz ve akran değerlendirme yöntemleri kullanıldığında en önemli sorun, bu kaynaklardan elde edilen puanların güvenirliği ve bu puanlara dayalı yapılan çıkarımların geçerliği olarak görülmektedir (Donnon, Mcllwrick ve Wololoschuk, 2013). Öğrencinin performansını etkileyen puanlayıcı kaynaklı faktörler puanlayıcı davranışları olarak adlandırılmaktadır (Farrokhi, Esfandiari ve Vaez Dalili, 2011). Bu bağlamda mevcut çalışmanın problem durumu, öz ve akran değerlendirmede hangi puanlayıcı davranışlarının ortaya çıktığı şeklinde belirlenmiştir.

Araştırmanın Amacı: Bu çalışmanın amacı, üniversite öğrencilerinin öz ve akran puanlama sürecinde hangi puanlayıcı davranışlarını sergilediklerini çok yüzeyli Rasch ölçme modeli aracılığıyla belirlemektir.

Araştırmanın Yöntemi: Araştırma öğretmen adaylarının hazırlamış oldukları araştırma önerilerinin puanlanması sürecinde göstermiş oldukları puanlayıcı davranışlarının

ortaya çıkarılmasını hedeflediği için var olan bir durumun betimlenmesinden dolayı betimsel türden bir nicel araştırma özelliği göstermektedir. Araştırmanın katılımcıları 2017-2018 eğitim ve öğretim yılında Ankara ilindeki bir vakıf üniversitenin eğitim fakültesi Rehberlik ve psikolojik danışmanlık programında yer alan bilimsel araştırma yöntemleri dersini alan öğrenciler arasından, çalışma kapsamında gönüllü olarak katılan 58 kişiden oluşmaktadır. Araştırma kapsamındaki veriler, araştırmacılar tarafından geliştirilen analitik dereceli puanlama anahtarı (ADPA) ile toplanmıştır. ADPA, herhangi bir bilimsel araştırma önerisini değerlendirmek amacıyla geliştirilmiştir. Öncelikle taslak olarak geliştirilen ölçme aracına yönelik olarak uzman görüşleri alınmıştır. Görüş ve öneriler doğrultusunda ölçme aracının son şekli verilmiştir. Buna göre, ölçme aracının ölçütleri; problem durumunun belirlenmesi, yöntem, bulgular ve sonuç/yorum olarak belirlenmiştir. ADPA'nın her bir ölçütü dörtlü bir derecelendirme (oldukça yetersiz "0", oldukça yeterli "3") kullanılarak puanlanmıştır. ADPA'dan elde edilen ölçümlerin geçerliği için AFA'i güvenirliği için ise McDonald ω katsayısı kullanılmıştır. Araştırmadaki verilerin analizinde; çok yüzeyli Rasch ölçme modeli kullanılmıştır. Analizler FACETS palet programı kullanılarak yapılmıştır. Analizinin bazı varsayımları bulunmaktadır. Bu varsayımların karşılanması analiz sonuçlarına dayalı yapılan çıkarımların geçerliğine hizmet etmektedir. İlk varsayım olarak tek boyutluluk incelenmiş olup veri toplama araçları kısmında ölçme aracının tek boyutluluğa sahip olduğu görülmüştür. Tek boyutluluğun sağlanması yerel bağımsızlığın da karşılandığının bir göstergesi olarak ele alınmış olup yerel bağımsızlık için herhangi bir işlem yapılmamıştır. Son olarak model veri uyumu incelenmiştir. Model veri uyumu için ±2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı toplam gözlem sayısının %5'inden fazla olmaması ve ±3 aralığının dışında kalan standartlaştırılmış artık değerlerin de toplam veri sayısının %1'inden fazla olmaması gerektiği belirtilmiştir (Linacre, 2017). Bu çalışmada toplam gözlem sayısı 2784 (58 x 12 x 4) olup, ±2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı 116 (%4.17) ve ±3 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı ise 28 (%1.ff01) olduğundan mevcut çalışma için model veri uyumunun sağlandığı görülmektedir.

Araştırmanın Bulguları: Araştırma kapsamında elde edilen bulgular incelendiğinde, kadın ve erkek puanlayıcıların benzer katılık ve cömertlik düzeylerine sahip oldukları bulunmuştur. Diğer yandan puanlayıcıların öz puanlamalarda daha cömert davranış sergiledikleri gözlemlenirken, akranlarını değerlendirme sürecinde ise daha katı davranış sergiledikleri gözlemlenmiştir. Ayrıca puanlayıcıların öz değerlendirmelerinin standart hatalarının akran puanlamalarına göre daha yüksek çıktığı bundan dolayı öz değerlendirmelerin güvenirliğinin daha düşük olduğu bulunmuştur. Puanlayıcıların grup düzeyinde merkeze yönelim davranışı sergilemedikleri fakat bireysel düzeyde 18 puanlayıcının merkeze yönelim davranışı sergilediği tespit edilmiştir. Diğer bir puanlayıcı davranışı olan farklılaşan katılık ve cömertlik durumları incelendiğinde, 696 olası etkileşiminin 69 tanesinin (%9.91) istatistiksel olarak anlamlı olduğu tespit edilmiştir. Grup çalışmalarının değerlendirilmesinde ortaya çıkan 69 bireysel anlamlı etkileşimin t-değerlerinin işaretine göre farklılaşan puanlayıcı katılığı ve cömertliği davranışı belirlenmektedir. Bu bağlamda mevcut çalışmada 69 anlamlı etkileşimden 14 tanesinin farklılaşan

puanlayıcı katılığı olduğu 55 tanesinin ise farklılaşan puanlayıcı cömertliği olduğu belirlenmiştir.

Araştırmanın Sonuç ve Önerileri: Araştırmada; puanlayıcıların cinsiyetlerine göre puanlamada katılık veya cömertlik davranışları anlamlı farklılık sergilememektedir. Bireysel olarak puanlayıcılardan 14'ü katılık ve cömertlik davranışı sergilemesi bu puanlayıcıların hem kadın hem de erkeklerden oluştuğunu göstermektedir. Bu nedenle katılık ve cömertliğe ilişkin puanlayıcı eğitim sürecinde her iki cinsiyet grubuna yönelik puanlama eğitimine alınması önerilebilir. Ayrıca puanlama sürecinde öz ve akran puanlayıcı türüne göre katılık ve cömertlikte farklılaşan davranışı gösterdikleri görülmüştür. Bu ise bireylerin kendi performanslarını veya akranların performanslarını değerlendirirken farklı davrandıklarını göstermektedir. Bu durum, puanlayıcıların daha objektif davranabilmesi için öz ve akran değerlendirme eğitimine yönelik çalışmalara daha fazla önem verilmesi gerektiğini göstermektedir. Özellikle öz puanlamalara yönelik, bir program dâhilinde çalışmalar yürütülebilir. Araştırmada grup bazında olmasa da bireysel bazda merkeze yönelme davranışı görülmüştür. Bu ise bireylerin puanlama kategorilerinin uç noktaları ile orta noktayı daha fazla tercih ettiği şeklinde açıklanabilir. Buradan da puanlayıcıların özellikle dereceli puanlama anahtarını daha dikkatli kullanımına yönelik çalışmalar üzerinde durulması önerilebilir.

Anahtar Kelimeler: Akran değerlendirme, Öz değerlendirme, Puanlayıcı yanlılığı, Yeni Yaklaşımlar, Çok Yüzeyli Rasch Modeli.