Analysis of the Rater Effects on the Scoring of Diagnostic Trees Prepared by Teacher Candidates with the Many-Facet Rasch Model

Funda Nalbantoğlu Yılmaz

Faculty of Education, Nevşehir Hacı Bektaş Veli University, Nevşehir, Turkey

fundan@nevsehir.edu.tr

Abstract

In the study, it was aimed to investigate the leniency/severity, bias and halo effect of the raters which were used in the scoring of the diagnostic tree prepared by the teacher candidates with the many-facet Rasch model. The research study group constitutes 24 teacher candidates who are taking measurement and evaluation lesson from the students of the faculty of teaching department of classroom teaching in a state university. Teacher candidates in the study team have formed two groups between themselves. Each group developed a diagnostic tree related with their field. Diagnostic trees prepared by teacher candidates were scored by a faculty member in the direction of the same criteria and 12 peers selected from each group. The effects of scoring were determined and the data were analyzed with the many-facet Rasch model. In the study, it was reached to the results that some raters rated severely on some criteria and groups than expected, and some raters showed halo effects on an individual-level.

Keywords: Diagnostic tree**,** Many-facet Rasch model, Leniency/severity, Bias, Halo effect

1. Introduction

Nowadays which the constructivist approach is dominant, it is necessary for teachers to have some competences such as using proper measurement and evaluation methods, developing appropriate tools, practice and using the results that are obtained. The effects of the measurement and evaluation lessons in the teacher training programs are high at the start of the ability to have these proficiencies of the teachers that are inaugurate.

Teacher candidates must have sufficient knowledge and skills in developing and implementing appropriate measurement tools in the measurement and evaluation lesson they take in their student life in order to be able to properly and accurately measure and evaluate when they start their careers.

Regarding the examination the self-efficacy of teachers on the measurement tools, it is observed that the teachers have low diagnostic tree levels and low frequency of use and that there is in-service training needs in this respect (Karamustafaoğlu, Çağlak, & Meseciler, 2012; Okur, 2008; Er-Nas & Çepni, 2009). In addition to these studies, Köklükaya, Öztuna-Kaplan and Sevinç (2014) found that the pre-service science teachers had a low level of diagnostic tree-building competence. Duran, Mıhladız and Ballıel (2013), in their studies about the proficiency levels of primary school teachers' measurement and evaluation methods that is examined, it was determined that the teachers see themselves very little sufficient in preparing diagnostic tree. Aydoğmuş and Coşkun-Keskin (2012) have determined in their research on the use of measurement and evaluation tools by social sciences teachers that a large majority of teachers in social sciences have never used a diagnostic tree. When the reasons why teachers did not use diagnostic tree were examined, it was determined that they did not have enough knowledge on this subject. In the literature, it was determined that the teachers had problems in using/developing diagnostic tree. In teacher training institutions, it is necessary to measure how much information is gained in teaching professional courses, as well as to measure how teacher candidates are able to use this knowledge in real life situations related to their profession. In this line of work, the process of diagnostic tree development by teacher candidates in an appropriate subject is considered.

Diagnostic tree is a technique in which a horizontal or vertical section is presented to students in the form of a tree branch with multiple correct false statements in relation to each other. The diagnostic tree consists of proposals that are go from the general to the specific to evaluate interrelated subjects, and the answer to be given to each question specifies the direction to be taken and the question to be answered in the next step (Başol, 2013).

In this study, in which teacher candidates examine the diagnostic tree development process, the rater effect in scoring diagnostic trees prepared by teacher candidates is an important effect to be investigated. Because a rater-based subjectivity may interfere to the individual performance scores (Eckes, 2009).

Raters play an important role in determining the performance of individuals in a situation. Rater errors can lead to errors on individual scores and can affect the scores negatively (Farrokhi & Esfandiari, 2011). For example, a lenient rater may rate an individual's performance better than it actually is, while a severe rater may rate worse (Eckes, 2009). In interpreting tasks or categories in the guideline of scoring, there may be a difference between the raters (Prieto & Nieto, 2014) in the rater leniency/severity, in the degrees of influencing from the general impression of the individual. The sources which were created this difference, are called rater effects. Rater effects are listed as rater leniency/severity, halo effect, central tendency, restriction of range, bias and inconsistency (Myford & Wolfe, 2003; Saal, Downey, & Lahey, 1980).

The rater leniency/severity is that one or more of the raters are in some situations depending on the individual performance rate more or less the performance of the individual than how it should be. The halo effect is a failing of the rater in distinguishing the different aspects of the student's performance and is a giving these features the similar rates (Engelhard, 1994). Central tendency effect is that the rater overrides the midpoint of the rating criteria (Myford & Wolfe, 2004), making clustered rating at the midpoint of the rating scale (Landy & Farr, 1983; Saal, Downey, & Lahey, 1980). The central tendency effect is the rater error that indicates how the ratings criteria categorization is used, which affects reliability and validity (Saal, Downey, & Lahey, 1980). It is a narrowing of the score ranging with the making scoring around a offcenter position on the rating scale of the rater (Eckes, 2015). Bias is to be in tendency to giving, to some individuals, lower or higher scores due to some effects of the rater (Engelhard, 1994). Rater effects influences the reliability and validity of performance scores. For this reason, in the situations that are used more than one rater, it is seen as a significant effect that needs to be searched.

The many-facet Rasch model has an important role in the observation of the rater effect in the performance determination process. The many-facet Rasch model is an extension of the one parametered Rasch model in item response theory. It provides rater leniency and severity in addition to inter-individual ability differences and substance abuse in the Rasch model (Linacre, 1989). We can observe the rater leniency/severity, rater consistency, functioning of scale and interaction between the variables with the many-facet Rasch model (Eckes, 2009). There are several advantages on using the many-facet Rasch model. One of these is the invariance feature. That is, the individual's ability is calculated independently from the items, tasks, group or rater. Together with this, item and rater criteria are calculated independently from the group (Eckes, 2009). Another advantage of the many-facet Rasch model is that it gives the rater effects (bias, halo effect, central tendency effect, etc.) to be investigated in situations where more than one rater is used.

In the literature, it has seen that there are studies examined the rater effets with the many-facet Rasch model in different situations (Bechger, Maris, & Hsiao, 2010; Eckes, 2005; Engelhard, 1994; Esfandiari, 2015; Farrokhi & Esfandiari, 2011; Farrokhi, Esfandiari, & Schaefer, 2012; Hung, Chen, & Chen, 2012; İlhan, 2015; Kassim, 2011; Myford & Wolfe, 2003; Myford & Wolfe, 2004). But in any case where raters are used, the investigation of the rater effects seems important for the reliability and validity of the rates obtained.

In the research, it was aimed to investigate the effects of the raters who took part in the rating of the diagnostic tree prepared by the teacher candidates in the research with the many-facet Rasch model. In the scope of the study, rater leniency/severity, bias and halo effects were considered as the rater effects. In this direction, it was searched answers to the following questions.

1. How is the variable map of the scoring of the diagnostic trees prepared by the teacher candidates by an instructor and 12 teacher candidates according to 11 criteria?
2. How is the leniency/severity of the raters used for the scoring of the diagnostic trees that the teacher candidates prepare for in a subject related to their field?

3. How is the analysis of bias related to the Rater x Group interaction?
4. How is the analysis of bias related to Rater x Task interaction?
5. Do the raters show halo effect?

2. Methodology

In the study, it was attempted to identify the status of the raters used in the scoring of the diagnostic tree-making skills of the teacher candidates in a given subject. For this reason, research is in the screening model.

2.1. Study Group

The study group consists of 24 teacher candidates who are taking measurement and evaluation lesson from the students of the education faculty classroom teacher department at Nevşehir Hacı Bektaş Veli university in the academic year 2015-2016. Teacher candidates in the study group have formed groups in double between themselves. Thus, the study was conducted over 12 study groups. A lecturer was used to rate the diagnostic tree prepared by the teacher candidates. In addition to the lecturer, each teacher candidate voluntarily selected from the groups formed within the scope of the study also rated the diagnostic trees.

2.2. Data Collection

The research was carried out in the course of measurement and evaluation of the department of classroom teaching at the education faculty of a state university in the 2015-2016 academic years.

Within the scope of the research, firstly, teacher candidates were presented with a diagnostic tree, examples were shown and discussed. Later on, teacher candidates were asked to prepare a diagnostic tree with their group mates on a topic determined by the curriculum. After the preparations, the form which will be used in the scoring of diagnostic tree that was prepared, was introduced to the class and a sample rating is made. Later, all of the diagnostic trees prepared by the groups were scored by a group of volunteer teacher candidates that were selected one a piece voluntarily from each group. The lecturer who was in the school rated the products of the teacher candidates through the same rating scale as the teacher candidates. In this case, each diagnostic tree was rated by 13 raters, including 12 teacher candidates and one lecturer.

A rating scale was used in rating of the diagnostic tree that was prepared. In the preparation of rating scale, performance is defined first and important aspects of the product to be revealed at the end of performance are determined. Literature search for diagnostic tree properties and development has been done. As a result of the screening, the related skill was defined by taking into account the features of diagnostic tree and the observable criteria were determined. The specified criteria were examined in terms of fitness, necessity, and clarity, by four measurement and evaluation experts and 12 teacher candidates who were selected from each study group voluntarily and who were trained in the diagnostic tree. In accordance with the opinions that were taken, some expressions that are similar to one another were corrected / removed. Then, according to the opinions of expert and teacher candidate for each article, the Lawshe (1975) content validity ratios (CVR) and content validity index (CVI) were calculated.

Table 1 shows the content validity index, the content validity ratio and rating scale calculated for each criteria according to the opinions of four measurement and evaluation experts and 12 teacher candidates.

Table 1. The content validity ratio and the content validity index related to the criteria

Criteria	Necessary	Necessary-Insufficient	Unnecssry.	CVR
1. The tree branches are schematized as a horizontal or vertical section.	16			1.00
2. The questions (true/false items) interrogate an interconnected information network.	14	1	1	0.75
3. Each expression placed in the branches is divided into two branches as true (T) / false (F).	16			1.00
4. The true/false items are in general-to-specific pattern.	13	1	2	0.625
5. The items are in the true (T) / false (F) format.	14	1	1	0.75
6. The items measure the correctness or the inaccuracy of just one statement.	15		1	0.875
7. There is no proposal which contains double negativity.	13	1	2	0,625
8. The true/false items are clear.	14	2		0.75
9. The true/false items have been expressed to be absolutely correct or absolutely wrong.	13	1	2	0.625
10. The exit points of the recognizable branch are specified.	15		1	0.875
11. Points related to exit points are calculated.	15	1		0.875
Content Validity Index (CVI)	0,80			
Content Validity Criteria (CVC)	0.42 <CVC <0.49			

When the content validity ratio is calculated, how the referees evaluate each expression is taken into account (Şencan, 2005). As shown in Table 1, the content validity ratios of the criterion expressions that all four of the experts and 12 teacher candidates reported as "necessary" were 1.00. However, Lawshe has calculated the minimum content validity ratio that each statement should have according to the number of experts used in the study, and suggested that the items which are below from this minimum should be subtracted from the rating form (Lawshe, 1975). When 15 referees are used in the research according to Lawshe's minimum content validity ratio table, the content validity ratio for each statement should be at least 0.49, and when 20 referees are used it should be minimum 0.42 (Lawshe, 1975). According to the content validity ratios that were given in the Table 1 in this respect, we can affirm that 11 criterion are in accordance with the relevant conceptual structure and that there is a concord between the referees regarding the criteria. Using all the criteria, the content validity index for the whole rating scale was calculated as 0.80. This value is greater than Lawshe's content validity criteria based on the number of referees in the criteria table. For this reason, the content validity of the rating form was found to be statistically significant.

The final form, prepared for the purpose of scoring of diagnostic tree prepared by teacher candidates, consists of 11 criteria. The criteria are rated as yes (2), partly (1) and no (0), depending on the status of having the relevant feature.

2.3. Data Analysis

In this study where the rater effects of the rating of the diagnostic tree that teacher candidates prepared in a subject in their field were observed, many-facet Rasch measurement model was applied. In the study, there are three facets; group, criteria, rater. Rater leniency/severity rates and fitness statistics, rater x group, rater x criteria interactions, halo effect are determined with many-facet Rasch analysis. The necessary analyzes were performed through the FACETS program (Version 3.64; Linacre, 2008). For the infit and outfit values which determine the degree of difference of the observed scores in the analysis from the expected scores, it was taken into consideration the 0.50 lower control limit suggested by the Wright & Linacre (1994), Linacre (2002, 2003, 2008) and the 1.50 upper control limit.

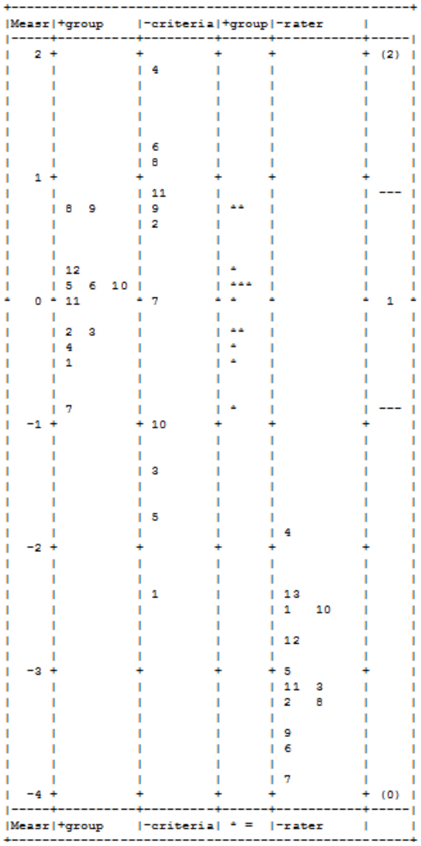
3. Results

The findings of the analysis are given below.

3.1. Model Data Fit

When the data fit the model, about 5% of standardized residuals are outside ±2, and about 1% are outside ±3 (Linacre, 2017). Out of the standard residual values of the total 1716 data obtained by rating of the diagnostic tree by the raters , 63 (3.67%) were outside ± 2 and 15 (0.87%) were outside ± 3. Accordingly, it leads to the interpretation that the analysis has provided the model data conformity.

The variable map is shown in Figure 1, which is obtained from examining the rating of the diagnostic tree through 11 criteria by a teacher and 12 teacher candidates with many-facet Rasch measurement model.



```
+---------------------------------------------------+
|Measr|+group     |-criteria|+group|-rater     |     |
|-----+-----------+---------+------+-----------+-----|
|   2 +           + 4       +      +           + (2) |
|     |           |         |      |           |     |
|     |           |         |      |           |     |
|     |           |         |      |           |     |
|     |           |         |      |           |     |
|     |           | 6       |      |           |     |
|     |           | 8       |      |           |     |
|   1 +           +         +      +           +     |
|     |           | 11      |      |           | --- |
|     | 8   9     | 9       | **   |           |     |
|     |           | 2       |      |           |     |
|     |           |         |      |           |     |
|     |           |         |      |           |     |
|     | 12        |         | *    |           |     |
|     | 5   6   10|         | ***  |           |     |
*   0 * 11        * 7       * *    *           *  1  *
|     |           |         |      |           |     |
|     | 2   3     |         | **   |           |     |
|     | 4         |         | *    |           |     |
|     | 1         |         | *    |           |     |
|     |           |         |      |           |     |
|     |           |         |      |           |     |
|     | 7         |         | *    |           | --- |
|  -1 +           + 10      +      +           +     |
|     |           |         |      |           |     |
|     |           |         |      |           |     |
|     |           | 3       |      |           |     |
|     |           |         |      |           |     |
|     |           |         |      |           |     |
|     |           | 5       |      |           |     |
|     |           |         |      | 4         |     |
|  -2 +           +         +      +           +     |
|     |           |         |      |           |     |
|     |           |         |      |           |     |
|     |           | 1       |      | 13        |     |
|     |           |         |      | 1    10   |     |
|     |           |         |      |           |     |
|     |           |         |      | 12        |     |
|     |           |         |      |           |     |
|  -3 +           +         +      + 5         +     |
|     |           |         |      | 11   3    |     |
|     |           |         |      | 2    8    |     |
|     |           |         |      |           |     |
|     |           |         |      | 9         |     |
|     |           |         |      | 6         |     |
|     |           |         |      |           |     |
|     |           |         |      | 7         |     |
|  -4 +           +         +      +           + (0) |
|-----+-----------+---------+------+-----------+-----|
|Measr|+group     |-criteria| * =  |-rater     |     |
+---------------------------------------------------+
```

Figure 1. Variable Map

In the variable map given in Figure 1, the columns for the groups show the ability of the groups to prepare diagnostic tree. The upper part of the column shows a high rate in terms of the relevant qualification and the lower part shows a low rate. Group 8 and 9 are the highest in terms of diagnostic tree making, while group 7 is the lowest.

The criteria column compares the measures used for rating the diagnostic tree prepared by the groups. The logit measures for the criteria range from +1.93 to -2.41. These values are spread over a wide area. This situation indicates that the criteria differ in terms of being able to do so. At the top of the column of measures it shows the difficulty of being able to make measurements comparing to the bottom. In this case, the 4th criterion indicates that the most difficult to prepare in diagnostic tree making for the teacher candidates, and the 1st criterion represents the easiest behavior.

The column for the rater in the data calibration map shows the rater leniency/severity. The upper part of the column shows the rater severity and the lower part shows the rater leniency. Raters' severity-leniency measures range from -1.86 to -3.86. Among the raters, the most severe is the 4th, while the most lenient is 7th.

3.2. Rater Analyzes

Findings such as leniency/severity, fitness statistics for the raters are given in Table 2.

Table 2. Rater Analysis Results

Observed	Observed	Logit	Model	Infit		Outfit		Raters
Score	Count	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	
217	132	-1.86	0.16	0.74	-1.9	0.48	-1.7	4
235	132	-2.42	0.19	0.81	-0.9	0.62	-0.7	13
236	132	-2.46	0.20	0.90	-0.4	2.20	2	10
238	132	-2.54	0.20	1.09	0.4	1.29	0.6	1
243	132	-2.77	0.22	0.94	-0.2	0.80	-0.2	12
247	132	-2.98	0.24	1.46	1.6	0.72	-0.2	5
250	132	-3.18	0.27	1.04	0.2	0.52	-0.6	3
250	132	-3.18	0.27	1.33	1.1	1.78	1.1	11
251	132	-3.25	0.28	0.98	0.0	1.37	0.7	2
251	132	-3.25	0.28	0.97	0.0	0.77	-0.1	8
254	132	-3.51	0.31	1.04	0.2	0.41	-0.6	9
255	132	-3.61	0.33	1.24	0.7	1.06	0.3	6
257	132	-3.86	0.37	1.66	1.4	0.79	0.0	7
244.9	132	-2.99	0,26	1.09	0.2	0.98	0.0	Mean
10.6	0.0	0.54	0.06	0.25	0.9	0.52	0.9	S. D.
RMSE (Model) :0.26		Separation: 1.88		Strata: 2.84		Reliability: 0.78		
Fixed (all same) chi-square: 69.9		d. f.= 12		p= 0.00				
Random (normal)) chi-square: 10		d. f.= 11		p= 0.53				

The rating leniency/severity measures ranged from -1.86 to -3.86 as given in Table 2. The separation index for the raters is calculated 2.84 through the formula (4G + 1) / 3 (G: separation ratio). In this context, it can be said that the rating severity is statistically divided into three classes. The separation index reliability is 0.78. In the case where the separation index reliability is close to zero, it emphasizes that there is no significant difference in the leniency/severity of the raters and the homogeneity of the raters (Engelhard, 2012). This is a desirable situation. At the same time, the reliability of the separation index is around 0.70, suggesting that there are significant differences between severe and lenient raters (Myford & Wolfe, 2004). The reliability coefficient of the rater was not as low as expected. We can claim at this point that there is a difference in the severity / leniency of the raters.

The hypothesis was rejected by the chi-square test ("There is no significant difference between raters in terms of leniency/severity ratings", $X^2= 69.9, df = 12, p < 0.01$). There are significant differences between the leniency/severity of the raters. To determine from which the raters these differences originate from, Myford and Wolfe (2004) suggested that the raters' position on the variable map, rater severity measures, and t-values of the raters should be examined.

When the distribution of the raters is examined in the data variable map in Figure 1, it is seen that the raters range from -1.86 to -3.86 logit values. The rater 7 is the most lenient rater according to rater logit values. The rater 4 is more severe than the other raters. The t-values are calculated to determine the different raters. For this, the logit measurements of each rater subtracted from the mean logit measure and the difference is divided by the standard errors (Çetin & Ilhan, 2017; Myford & Wolfe, 2004). The t-values calculated for each rater were compared with the critical t-value of 3.055 at the significant level 0.01 in 12 degrees of freedom. The t-values for the raters are given in Table 3.

Table 3. Rater t-Test Results

Rater	t-value	Rater	t-value
P4	7.0625	P11	-0.704
P13	3	P2	-0.93
P10	2.65	P8	-0.93
P1	2.25	P9	-1.677
P12	1	P6	-1.878
P5	0.042	P7	-2.351
P3	-0.704		

With reference to the t-values given in Table 3, we can state tvalue> tcritical value only for the rater 4. In this case, it can be said that the difference between the raters is mostly derived from the 4th rater.

When the random effect hypothesis in Table 2 is tested with chi-square (X^2*= 10, df = 11, p > .05*) rate, it can be said that the distributions are normal, and the data are consistent with the model. Fitness statistics indicate the level of consistency of the raters when determining the abilities of the individual (Lunz, Stahl, Wright, & Linacre, 1989). As shown in Table 2, the infit values of the raters range from 0.74 to 1.66. The outfit values range from 0.41 to 2.20. Acceptable infit and outfit values are between 0.50 and 1.50 according to Linacre (2002). In this case, it can be said that the infit values of the other raters except 7th rater are acceptable. 1.5 and above, the fitness statistics point to the inconsistency, unpredictability of the scoring points. The fitness statistics of 0.5 and below indicates that the differences in the points are not sufficient to limit the range (Lunz, Stahl, Wright & Linacre, 1989; Yan, 2014). Referring to the outfit values of the raters in Table 2, it is determined that the outfit statistics of 4th and 9th raters are lower than the lower control limit and the outfit statistics of raters 10 and 11 are higher than the upper control limit. In this case, it can be said that the difference between raters of 4th and 9th raters is not sufficient. At the same time, this also indicates that there may be a restriction of ranges for these raters. It can also be said that in the rating of raters 10 and 11, there are inconsistencies.

3.3. Bias (Interaction) Analysis

Raters x Group interactions have been examined to determine whether raters behave severely or leniently towards the groups in rating. In Table 4, Rater x Group interaction is given.

Table 4. Rater x Group Bias Analysis

Observed Score	Exp. Score	Obs-Exp Average	Bias Measure	Model S.E	z-Score	Rater	Group
18	20.94	-0.27	-1.39	.55	-2.53	13	9
15	18.96	-0.36	-1.07	.48	-2.23	1	1

Rater x Group bias analysis shows whether a particular group of raters rated with a certain leniency/severity (Myford & Wolfe, 2003, 2004). The bias size is greater than zero, which indicates that the observed rate is greater than the expected rate. When the estimated z-score for Rater x Group bias is greater than +2 or less than -2 (p <.05 at the significant level), it results from bias (Myford & Wolfe, 2004).

According to Table 4, there are 2 significant Rater x Group interactions out of 156 data belonging to Rater x Group interaction (13 raters x 12 groups). This Rater x Group interactions show negative bias. In the study, it was determined that 2 raters (raters 1 and 13) rated more severely than expected in 2 groups (1 and 9 groups).

The Rater x Criteria interactions were examined to determine whether the raters rated the same severity-leniency. Rater x Criteria bias analyzes are given in Table 5.

Table 5. Rater x Criteria Bias Analysis

Observed Score	Exp. Score	Obs-Exp Average	Bias Measure	Model S.E.	z-Score	Rater	Criteria
14	19.76	-0.48	-1.02	0.38	-2.68	13	8
13	19.29	-0.52	-1.06	0.38	-2.79	13	6
20	23.01	-0.25	-1.37	0.51	-2.69	12	7
17	22.81	-0.48	-1.81	0.41	-4.42	11	2
23	23.89	-0.07	-2.18	0.98	-2.23	11	5
22	23.66	-0.14	-1.73	0.70	-2.47	10	3
14	19.12	-0.43	-0.87	0.38	-2.29	4	9
18	22.01	-0.33	-1.13	0.43	-2.63	3	8
21	23.59	-0.22	-1.93	0.58	-3.33	1	10

According to the results of the bias analysis of the Rater x Criteria interaction (13 raters x 11 criteria) given in Table 5, 9 out of 143 data show significant bias. The first rater rated the 10th criterion, the third rater rated the eighth, and the fourth rater rated the ninth severely. While the rater 10 rated the 1st criterion severely, the rater 11 was 2nd and 5th, the rater 12 was 7th, the rater 13 was the 8th and the 6th criteria severely. Raters 13, 12, 11, 10, 4, 3 and 1 tend to give lower rates to certain criteria than expected.

3.4. Halo Effect

The halo effect is that the thing which rater fails to distinguish the different aspects of the student's performance and gives them similar rates (Engelhard, 1994). Two types of halo effect can be mentioned at individual and group-level (Esfandiari, 2015, Myford & Wolfe, 2004). In the halo effect at the group-level, each rater tends to rate the same rating each student on different criteria. For this reason the raters cannot distinguish the criteria (Esfandiari, 2015). For halo effect at the group level, information about the criteria used in rating should be examined. For halo effect at the individual-level, the raters are to be evaluated for infit and outfit values (Myford & Wolfe, 2004).

In determining the halo effect at the group-level, the separation statistics of the criteria used in the study were examined, as suggested by Myford and Wolfe (2004). The separation statistics for the criteria are given in Table 6.

Table 6. Separation Statistics for Criteria

Observed	Observed		Model	Infit		Outfit		Criteria
Score	Count	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	
289.5	156	0.00	0.37	1.05	0.2	0.98	0.2	Mean
21.3	0.0	1.34	0.27	0.20	0.8	0.19	0.5	S.D.

RMSE (Model) : 0.46 Separation: 2.73 Strata: 3.98 Reliability: 0.88

Fixed (all same) chi-square: 132.3 d.f.= 10 p= 0.00

The rater showing halo effect will rate almost all the criteria with the same rate, and there will be apparent similarity between the criteria. Statistically insignificant chi-square values of the criteria, low separation index, low separation index reliability indicate the criteria similarities. This may suggest that the raters cannot distinguish the criteria, and whether or not they have the halo effect (Myford & Wolfe, 2004). As given in Table 6, the chi-square value of the criteria is statistically significant and indicates that there are differences between at least two measurable cases ($X^2= 132.3, df = 10, p< 0.01$). The separation index (strata) is 3.98, and the separation index reliability is 0.88. We can affirm that the criteria used in this direction gives reliable results in distinguishing the diagnostic tree prepared by groups and according to the performances of individuals. This event indicates that the raters do not tend to rate the criteria in the same rating and statistically do not suggest halo effect at group-level.

Myford and Wolfe (2004) suggested examining the fitness statistics of the rater to determine the halo effect at the individual-level. In this study, Myford and Wolfe (2004) used a measure of fitness statistics that is significantly smaller than 1 when there is little difference between the levels of difficulty of the items / attributes, and significantly greater than 1 when the differences between the difficulty levels of the items are great. Ideally, infit and outfit values should be close to 1. Infit values below 1 may indicate that the rater tends to give the same rate, and if they are above 1, they may show randomness in the rates (Linacre, 2002). Referring to the variable map in Fig. 1 and the separation statistics for the criteria in Table 6, we can affirm that there are differences between the cases in which the criteria in the study can be made. For this reason, it has been taken into account that the fitness statistics of the raters show excessive deviations from 1 to determine halo effect at the individual-level. According to the rater fitness statistics in this line of work, 10th (outfit: 2.20), 11th (outfit: 1.78) and 7th (infit: 1.66) raters may be a matter of halo effect.

4. Conclusion and Discussion

In the study, it was aimed to investigate the raters leniency/severity, Rater x Group, Rater x Criteria interactions and the halo effect used in the rating of the diagnostic tree prepared by the groups with the many-facet Rasch model. Analyzes made for this purpose have resulted in significant differences between the leniency/severity of the raters. The most lenient of the raters are 7th, and the 4th rater is the most severe. However, it was determined that the differences between rating of the 4th and 9th raters were not sufficient, and the rating of 7th, 10th and 11th raters were inconsistent.

It has been determined that some raters (13, 12, 11, 10, 4, 3, and 1) tend to rate lower than expected when examining Rater x Criteria interactions. These raters are teacher candidates selected from the groups. Teacher candidates may have lower rates on some of the criteria than expected, which may suggest a bias in terms of their intelligibility and rater behavior. While the criteria were being prepared, opinions were gathered from both experts and teacher candidates that the criteria were intelligible. However, in the analyzes, it was determined that the criteria used for the reliability of the separation index and separation index for the criteria provided reliable results in distinguishing the individuals according to their performance. In addition, it has been determined in the study conducted that the fitness statistics for the criteria have acceptable usage characteristics. In this respect, we can assume that the Rater x Criteria bias may be derived from the professional inexperience of the teacher candidates in terms of rating rather than the criteria.

Also, in the Rater x Group bias examination, it is determined that rater 1 rated more severely than normal in group 1 and rater 13 rated more severely than normal in group 9. In the study, a teacher candidate (peer) was selected from each group and these peers were used as a rater together with the faculty member. Raters 1 and 13 are peers selected from groups. In this respect, it can be said that peer raters 1 and 13 are severely biased in the rating of the diagnostic tree prepared by the groups. In a study conducted by Farrokhi, Esfandiari and Schaefer (2012) as well, it was determined that peers in general tend to rate lower on some students than expected. More severe or lenient rating of the raters may be influenced by various factors such as professional experience, personality traits, attitudes, demographic characteristics, workload (Eckes, 2009). As a result of the research, it has been determined that the peers used as raters did more severely rating than expected. We can suggest that this results from the inexperience of peers in professional and rating practice and peer influence.

There was no halo effect at the group-level in the study. However, individual-level examinations suggest a halo effect for the raters 7, 10 and 11 (peers). Twelve peers selected from the groups in the study were used as raters beside the lecturer. As in this study, in other studies where peers within the class were used as raters, halo effects were stated at individual-levels of peers (Engelhard, 1994, Esfandiari, 2015; Farrokhi & Esfandiari, 2011). Myford and Wolfe (2003) suggested that the distinction between the criteria for reducing the halo effect should be understood by the raters. In order to achieve this, it is necessary to make more exercise of rating to the teacher candidates used in the research. However, in the literature studies, it is suggested that a training related to rating, rating scale and criteria should be provided in order to reduce the rater effects (Farrokhi, Esfandiari, & Schaefer, 2012; Ilhan & Çetin, 2014; Myford & Wolfe, 2003). A brief training on the criteria was given before the rating in the study. It has been determined that some peer raters rated more severely than expected. From the results obtained, it is understood that there is a need for further training and rating examples of criteria and rates.

References

Aydoğmuş, A., & Coşkun Keskin, S. (2012). The situation of social studies teachers' using of process-oriented assessment and evaluation instruments: A sample of Istanbul. *Mersin University Journal of the Faculty of Education*, 8(2), 110-123.

Başol, G. (2013). *Eğitimde ölçme ve değerlendirme.* Pegem Akademi, Ankara.

Bechger, T.M., Maris, G., & Hsiao, Y.P. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement*, 34(8), 607-619.

Çetin, B., & İlhan, 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.

Duran, M., Mıhladız, G., & Ballıel, B. (2013). The competency level of elementary school teachers' towards the alternative assessment methods**.** *Mehmet Akif Ersoy University Journal of The Institute of Educational Sciences*, 2(2), 26-37.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2(3), 197-221.

Eckes, T. (2009). *Many-facet Rasch measurement.* In S. Takala (Ed.), Reference supplement to the manual for relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Eckes, T. (2015). *Introduction to many facet Rasch measurement*. Peter Lang Edition, Germany.

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

Esfandiari, R. (2015). Rater errors among peer-assessors: Applying the many-facet Rasch measurement model. *Iranian Journal of Applied Linguistics (IJAL)*, 18(2), 77-107.

Farrokhi, F., & Esfandiari, R. (2011). *A* Many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies*, 1(11), 1531-1540. doi:10.4304/tpls.1.11.1531-1540.

Farrokhi, F., Esfandiari, R., & Vaez Dalili, M. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal*, 15, 70-77.

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), 79-102.*

Hung, S.P., Chen, P.S., & Chen, H.C. (2012). Improving creativity performance assessment: A rater effect examination with many facet Rasch model. *Creativity Research Journal*, 24(4), 345-357.

İlhan, M., & Çetin, B. (2014). Rater training as a means of decreasing interfering rater effects related to performance assessment. *Journal of European Education*, 4(2), 29-38.

İlhan, M. (2015). The identificat,on of rater efects on open ended math questions rated through standart rubrics and rubrics based on the solo taxonomy in reference to the many facet Rasch model. Published Doctorate Thesis. Gaziantep University, Educational Sciences Institute, Gaziantep.

Karamustafaoğlu, S., Çağlak, A. & Meşeci, B. (2012). Self-efficiency of primary school teachers related to the alternative testing and assessment tools. *Amasya Education Journal*, 1(2), 167-179.

Kassim, N. L. A. (2011) Judging behaviour and rater errors: an application of the many-facet rasch model. *GEMA: Online Journal of Language Studies*, 11 (3), 179-197.

Köklükaya, A. N., Öztuna-Kaplan, A., & Sevinç, V. (2014). Determination of Pre-service Science Teachers' self-efficacy perceptions and efficacy levels about the diagnostic branched tree technique. *Journal of Turkish Science Education*, 11(1), 63-74.

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. San Diego, CA: Academic Press.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563–575.

Linacre, J. M. 1989. *Many-facet Rasch measurement*. Chicago: MESA Press

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16(2)*, 878.

Linacre, J. M. (2008). *Facets Rasch model computer program* [Software manual]. Chicago: Winsteps.com.

Linacre, J. M. (2010). *Facets Rasch measurement computer program,* version 3.66.2. Chicago, IL: Winsteps.com.

Linacre, J. M. (2017). *A user's quide to FACETS. Rasch-model computer programs*. Chicago IL.

Lunz, M. E., Stahl, J. A.. Wright, B. D., & Linacre, J. M. (1989). Variation among examiners and protocols on oral examinations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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.

Okur, M. (2008). Determination of 4. and 5. class primary teachers' opinions of using alternative assessment and evaluation techniques at science and technology course. Unpublished Master Thesis. Zonguldak Karaelmas University, Social Sciences Institute, Zonguldak.

Özdemir, S. M. (2010). Elementary teacher competencies and ınservice training needs in alternative measurement and assessment tools. *The Journal of Turkish Educational Sciences*, 8(4), 787-816.

Prieto, G., & Nieto, E. (2014). Analysis of rater severity on written expression exam using many faceted Rasch measurement. *Psicológica, 35*, 385-397.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413-428.

Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenirlik ve geçerlilik.* Seçkin Yayıncılık, Ankara.

Şenel Çoruhlu, T., Er Nas, S. & Çepni, S. (2009). Problems facing science and technology teachers using alternative assessment tecnics: Trabzon sample. *YYU Journal of Education Faculty*, 1(1), 122-141.

Wright, B. D., & Linacre, J. M. 1994. Reasonable mean-square fit values. *Rasch Measurement Transactions,* 8, 370.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: Amixed-methods approach. *Language Testing*, 31(4) 501-527.