Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.7, No.32, 2016

An Analysis of Peer Assessment through Many Facet Rasch Model

Melek Gülşah Şahin
Faculty of Education, Gazi University, Ankara, Turkey
melekgulsah@gmail.com

Gülşen Taşdelen Teker
Faculty of Education, Sakarya University, Ankara, Turkey
gtasdelen@sakarya.edu.tr

Neşe Güler
Faculty of Education, Sakarya University, Sakarya, Turkey
gnguler@gmail.com

Abstract
This study analyses peer assessment through many facet Rasch model (MFRM). The research was performed with 91 undergraduate students and with lecturer teaching the course. The research data were collected with holistic rubric employed by 6 peers and the lecturer in rating the projects prepared by 85 students taking the course. This study analyses raters, measurements for students who are rated, criteria used in rating and extent to which rubrics fulfil their function. Moreover, it also investigates effects of peers' levels of achievement on the process. In consequence, it was found that raters differed in the levels of strictness and generosity in rating, and that students were distinguished adequately in terms of the property measured. Besides, a very high level of reliability value was estimated in relation to the criteria in the study. This was interpreted as that they functioned in a reliable way in distinguishing between students' performances. It was found in the analyses of achievement levels of peers taking part in peer assessment that ratings made by students with high levels of achievement differed significantly from those made by students with medium or low level of achievement. Finally, the views about peer assessment were generally positive.
Keywords: peer assessment, many facet Rasch model, levels of peer achievement, rubric

1. Introduction

Significant changes occurred today owing to rapid development in information and technology. Changes occurred in the components of education as a result of adopting constructivist approach in particular in the field of education. Thus, an original and practice-based conceptual framework was formed in many topics ranging from the regulation of learning activities to the configuration of assessment process (Yurdabakan, 2011). The effects of changes emerging in education on especially the field of measurement and evaluation are also inevitable. Alternative methods of assessment have gained importance with the adoption of constructivist approach.

Identifying students' learning difficulties, monitoring their levels of learning continuously, and thus making the necessary improvements has become possible in alternative approaches of assessment (Çepni, 2007). Here, as different from traditional approaches, students are not only passive receivers of knowledge. Students' gaining a critical, creative and problem solving perspective of activities and issues with the development of higher order cognitive skills is the key point in this approach (Kutlu, Yıldırım & Bilican, 2009).

The importance of peer assessment increased with the adoption of alternative approaches of assessment. In peer assessment, students evaluate their peer or peers in a group (Falchikov, 1995; Freeman, 1995). The number of individuals having the responsibility can be one or more in this assessment. Falchikov (1994) suggests that the most important function of peer assessment in education is to provide detailed peer feedback (Falchikov & Goldfinch, 2000). Peer assessment is included in process evaluation to give feedback, and in level determination to determine achievement (Bostock, 2009). In addition to the above mention advantages, peer assessment has several advantages, some of which are described below.

- Students take on duties as individuals with responsibility in the process of assessment, not as the passive receivers of feedback (Sandvoll, 2014).
- Peer assessment raises the quality of learning process, improves the experience of critical thinking, and increases learner autonomy (Falchikov & Goldfinch, 2000; Pope, 2001; Noonan & Duncan, 2005; Lee, 2015).
- Peer assessment also affects metacognitive skills directly (Topping, 2005), and it makes the learning process possible.
- With peer assessment, individuals will have the opportunity to inquire the process of their learning (Tsai, 2000; Chou & Tsai, 2002).
- It was found in research that students became advantageous in learning by obtaining their peers' ideas

and views (Landry, Jacobs & Newton, 2015).

- Peer assessment practice performed in group work can also increase individuals' success because such practice increases the responsibility to contribute to group work (Yurdabakan & Cihanoğlu, 2009).
- Peer assessment is generally useful, reliable and valid practice (Falchikov, 1995).

Topping (2009) points out that peer assessment can be applied in differing fields and subjects. Race (1999) recommends using peer assessment in evaluating presentations, written plans, portfolios and exhibitions (Langan & Wheater, 2003). It is appropriate to use grading instruments or checklists in assessment. The criteria here can be set prior to assessment as well as during assessment according to needs (Falchikov, 2006). Yet, it is important that the criteria be clear and understandable so that an assessment serves to the purpose.

A review of literature demonstrates that peer assessment has restrictions beside advantages. Falchikov (1995) states that some of the students can hesitate to take on the responsibility of assessing their peers or that students with low achievement may not accept feedback from their peers. Topping (1998) says that students may tend to give higher marks in some cases than teachers. Dancer and Dancer (1992) points out that peers may have the tendency to perform assessment on the basis of similarities, race and friendship and thus to give higher scores unless they are offered detailed training (Yurdabakan, 2011). Ellington (1997) claims that failure to monitor peer assessment sufficiently can result in personal conflicts, prejudice and the formation of an atmosphere of competition between students – which is contradictory to the purpose.

It is possible to cope with the restrictions by trying to attain reliability and validity in peer assessment. Agreement between peer assessments can be analysed for reliability analysis, and the appropriateness of students' assessment according to the standards set by the teacher can be analysed for validity (Falchikov & Goldfinch, 2000).

Many studies are available in national and international literature in relation to peer assessment. Yurdabakan and Oğlun (2011) found in their research they conducted with elementary school students that self-assessment and peer assessment had positive effects on learning and on metacognitive knowledge levels. Bozkurt and Demir (2013) consult to elementary school fifth grade students' and their teachers' views of peer assessment. Accordingly, the teachers stated that peer assessment helped students to be objective and to gain experience in assessment and that it enabled them to revise knowledge- which was positive. As the negative side of peer assessment, they stated that it was difficult to be objective and that it was necessary to have a great number of activities. Besides, it was also found that the students who scored higher in peer assessment had more positive views than the other students. Yurdabakan (2012) analysed the effects of peer assessment activities on self-assessment skills. High correlations were found between activities and self-assessment in the experimental group in the study which was conducted in an experimental manner, and it was also observed that peer assessment was a good predictor of self-assessment. Falchikov (1995) state that there are correlations between teachers' assessment and peer assessment. The researcher also points out that an instrument of assessment has been developed for peer assessment feedback and students employ reflection, analysis and critical thinking skills here.

This study analyses the variables for peer assessment through many facet Rasch model (MFRM). Studies employing the MFRM in peer assessment research are also available in the literature. Aryadoust (2015) used not only Many-facet Rasch measurement but also correlation, and variance analysis. The results of his study show that the student raters, tutor, items, and rating scales achieved quite high psychometric quality. Besides these result, there were low to medium correlations between self, peer, and tutor assessments, and there was a significant difference between them. Karakaya (2015) compare self, peer and instructor assessment by using MFRM. As a result of the study, it was found that self-assessments were the most lenient while peer assessments were the severest. Moreover, statistically significant difference among the raters was found. Saito (2008) examined the training effect on peer assessment by using Many-facet Rasch measurement. There were control and treatment groups in the study. Both of the groups received instruction on skill aspects, but only the treatment group received long training on how to rate performances. The results show no significant differences between the groups.

1.1 The Significance and Purpose of the Study

This study primarily aims to examine the concepts of reliability and validity in relation to peer assessment. In line with this purpose - with ratings made by peer raters and ratings made by the teacher - the following are examined in details with MFRM:

- Raters operating,
- Measurements for the students rated,
- Criteria used in rating,
- The extent to which the levels used in rating key perform their function.

The fact that there are only limited number of studies performed on the basis of the MFRM in the literature (Aryadoust, 2015; Saito, 2008) makes this current study important. Thus, it was assured in this study that analysis for each facet mentioned above was done separately, and that detailed information was obtained for the facets available in the study. In addition to that, the views held by students performing peer assessment were obtained

through a form of open-ended questions. In this way, positive/ negative views in relation to the process were reached by means of answers given by peers taking on the role of raters. It is thought that the data obtained will be useful to those who are to do peer assessment practice in the future since they will give an idea about the practice. Besides, an attempt is also made in this study to determine the effects of achievement levels (low-medium-high) of students assigned to assess their peers on assessment process, which is a rarely studied issue in the literature. The fact that no studies were encountered in the literature in relation to this makes this study more important. Thus, in accordance with this second purpose, measurement reports prepared with MFRM in relation to ratings made by students according to their levels of achievement without considering teachers' ratings were analyzed, and whether or not they had any effects was determined. Based on this data, it is important in obtaining more reliable measurement results to inform teachers who are to use peer assessment in their classes that there is another factor they should take into consideration.

2. Method

2.1 Type of the Study

This is a descriptive survey. The properties of individuals, groups or of physical environments are summarized in studies using descriptive survey (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz, & Demirel, 2010).

2.2 Study Group

The study group was composed of 91 second year students attending the English Language Teaching department of a state university and taking the course of Scientific Research Techniques in 2014-2015 academic year.

2.3 Research Data

Six of the 91 undergraduate students included in the study group were included in the peer assessment process, and the following steps were taken in determining the students' function in the process:

- Firstly, all students were divided into three groups as low, medium and high achievement levels by taking into consideration the mid-term results and their cumulative grade point averages.
- Then two students were chosen randomly from each group. Each student selected were asked whether or not they wished to take part in the process of peer assessment, and thus six students of three different achievement levels were selected on the basis of volunteering.
- The remaining 85 students presented the projects for the course of Scientific Research Techniques to their friends who were assigned to assess their peers and to their teacher at the end of the semester.

Having determined the students to perform peer assessment, the nine-item integrated rubric shown in Table 1 was prepared by the lecturer of the course for use in assessment. The items were arranged in 5-pointed Likert type between "not appropriate" corresponding to 1 and "completely appropriate" corresponding to 5. Expert opinion was obtained from two experts of measurement and evaluation for the rubric developed, and it was used in the research after it was approved by the experts in terms of usability and clarity of statements.

Table 1. The rubric used for peer assessment

Criteria	completely inappropriate ⟹				completely appropriate
	1	**2**	**3**	**4**	**5**
1. The problem of the research was explained accurately.					
2. The sub-problems were explained obviously and intelligible way.					
3. Assumptions were denoted accurately.					
4. Limitations were explained appropriately.					
5. The type of the research was defined accurately.					
6. The study group was specified appropriately.					
7. The variables were defined accurately.					
8. The data collection tool was appropriate.					
9. The results of the research were explained obviously.					

So that the process of peer assessment could function more efficiently, the students to perform peer assessment were offered information on the criteria of measurement. In other words, the points to be considered in measuring the variables within the scope of the course with end-of-the-term projects were explained them in details. Besides, the importance of peer assessment was emphasized, and the students were reminded to be unbiased in their rating.

After rating was finished, the peer raters' views of the process were obtained with a form containing open-ended questions. Hence, the target was to reach the peer raters' positive and negative feedback.

2.4 Data Analysis

The quantitative data obtained in the study were analysed with MFRM using the FACETS Programme, which was developed by Linacre (2007).

Georg Rasch (1960), a German mathematician, established a mathematical model transforming the observation results at the level of ordinal scale into linear measurements. Although the model is in the form of a logistic regression model, each individual is a separate parameter for the items. Such that the model is similar to a regression in which each individual and each item is applied to a dummy variable as a coefficient and which is thus parameterized. The dummy variable here is "1" if the individual or the item is included in the desired observation (for instance if the individual answers correctly), and it is "0" otherwise (Quoted by Linacre, 2007). The Rasch model for the dichotomous data where the individuals and the items are included is:

$$\log\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i \qquad (1)$$

Whereas the many facet Rasch Model in which graded rating is performed and the diverse sources of variable such as raters apart from the individuals and the items are included was developed by Linacre as:

$$\ln\left(\frac{P_{nijk}}{P_{nij(k-1)}}\right) = B_n - D_i - C_j - F_k \qquad (2)$$

(Linacre, 1989; Akın and Baştürk, 2012). In this equation,

P_{nijk} = expresses the probability of individuals receiving k scores from J raters,

$P_{nij(k-1)}$ = expresses probability of individuals receiving k-1 scores from J raters,

B_n = expresses the individual's level of ability,

D_i = expresses the difficulty level of the item i.

C_j = expresses the rater's level of strictness/generosity in rating,

F_k = expresses the difficulty level in passing from the k-1 into k in rating,

(Linacre, 1989; Linacre and Wright, 2002; Looney, 2012). Thus, the many facet Rasch measurement model transforms the measurement results obtained for each source of variability into a real interval scale known as the "Logit Scale". In addition to the estimation of logits for each facet, the Rasch Model also enables the calculation of the significance of the differences between the facets under consideration (such as individuals' abilities and the difficulty level of the items) (Revesz, 2012).

Especially in cases of measurement where the sources of variability apart from individuals and items are available, the many facet Rasch model (MFRM) has certain advantages over the classical test theory; namely, (a) The MFRM helps to remove the bias problems arising in the direction of scores close to the average and against extreme scores (Bond and Fox, 2001; Quoted by Revesz, 2012). The differences between scores close to the average results is small in analyses performed with raw scores (for instance, in cases where the average is 50, the difference is between 48 and 55), and it is supposed that the same difference is available for the extreme scores (for instance 88-95). In other words, it is thought that the differences between scores correspond to real differences; indeed, individuals can only be ordered according to their levels of ability on the basis of the raw scores, and the real differences between individuals' abilities cannot be known. With the mathematical model on which the Rasch analysis is based, the log-odds are found for the raw data, they are transformed into the interval scale, and thus the problem is dealt with. (b) The estimations for each facet are made independently of the other facets in the MFRM model. For instance the estimation of individuals' ability levels is independent of the raters' levels of severity/generosity in rating. (c) If there is any waste of data, the analysis performed with the raw scores can yield false results whereas the wasted data are skipped in Rasch analysis, and the analyses are done only on the basis of observable data. (d) A separate reliability coefficient is found for each facet in the MFRM model, and the unexpected situations in each facet can be revealed one by one through fit tests (that is, the INFIT and the OUTFIT tests). (e)It forms a joint scale (a logit scale) in which the parameter estimations for each facet are included, and thus it provides great flexibility in interpreting the facets (Baştürk and Işıkoğlu, 2008; Güler and Gelbal, 2010; Kim, Park and Kang, 2012; Linacre, 1989).

3. Findings and Discussion

The findings for the raters participated in the research, the measurements for the students rated, the criteria used in ratings, the extent to which the levels in the rubric performed their function, and the effects of the peer raters' achievement levels on the process are presented respectively below. Besides, the responses given by the students performing peer assessment to the open-ended questions are also given in this part of the study.

The logit map in which the scores obtained by students whose performances for peer assessment process are measured, raters' strictness/generosity levels and the criteria for assessment are included is shown in Figure 1. The scale levels for the logit map are in the -3 and +3 range. The second column shows the score distribution for

the students whose performances are measured in the assessment, the distribution is from the top to the bottom- that is to say, from the students with the highest achievement to the ones with the lowest achievement. Accordingly, the student with the highest achievement is number 80 while the one with the lowest achievement is number 64. The third column of the logit map shows the criteria according to which students are assessed, and criterion 7 (describing the variables in the research accurately) shown at the top is the most difficult criterion whereas criterion 9 (stating the findings and conclusions clearly and in an understandable way) is the easiest. Column four, on the other hand, shows the distribution of the raters.

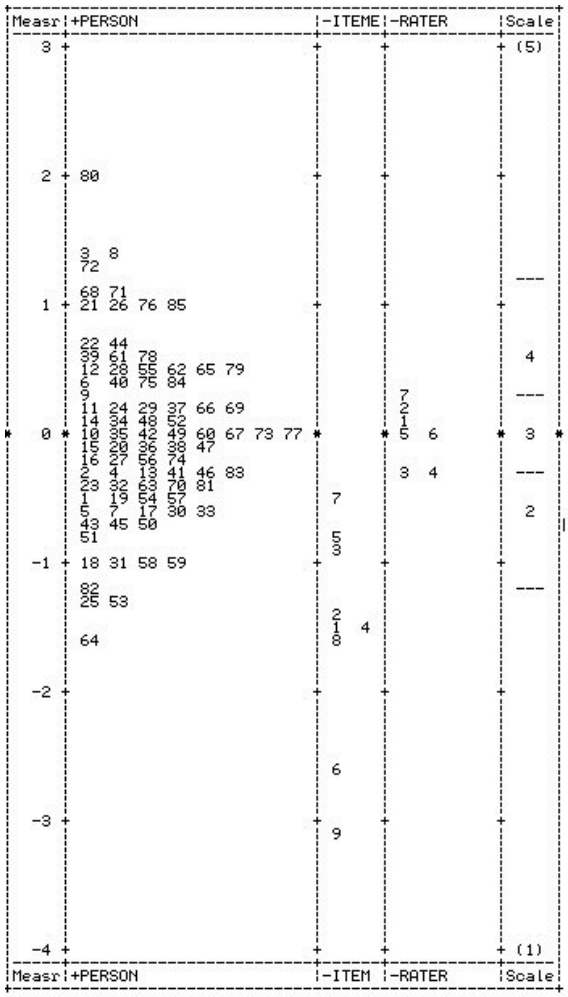

Figure 1. The Logit Map

Accordingly, rater 7 was the strictest rater (also the student with the highest achievement) but raters 3 and 4 (the students with low and medium achievement levels respectively) are the most generous raters. The final column shows the distribution of degrees of assessment (scoring between 1 and 5). Logit maps having all these data enable us to examine the facets visually on the same table.

The results for measurement reports in relation to the raters performing peer assessment are shown in Table 2.

Table 2. Raters' Measurement Report

Raters	Measure	Standard error	Infit	Outfit
R7 (high)	0.29	0.04	1.06	0.99
R2 (low)	0.17	0.04	1.13	1.10
R1 (instructor)	0.13	0.04	1.11	1.05
R5 (high)	0.00	0.05	0.93	0.87
R6 (medium)	-0.03	0.05	0.92	0.88
R4 (medium)	-0.27	0.05	0.98	0.79
R3 (low)	-0.29	0.05	1.16	1.21
Mean	0.00	0.05	1.04	0.99
Standard deviation	0.20	0.00	0.09	0.14
Reliability = 0.95 Seperation = 4.53 chi-square=124.3 df = 6 p = 0.00				

As is clear from Table 2, reliability and discriminant indices are 0.95 and 4.53 respectively according to the measurement reports. The quite high reliability value is the reliability value of the raters' differences, not of similarities (İlhan, 2016). In addition to that, the high discriminant index also represents the level of differences between ratings performed by the raters. When raters rate in complete consistency, discriminant index takes on the value of zero (Güler, 2014). Briefly, when these two figures are taken into consideration, it may be concluded that the raters' strictness and generosity levels in rating have differed. It is also observed accordingly that rater R7 is the strictest rater while rater R3 is the most generous rater. Yet, the fact that the infit and outfit statistics are in the 0.6-1.4 range, which is an acceptable interval, indicates that raters' scores fit the model (Nakamura, 2002; Wrigt and Linacre, 1994; cited in Semerci, 2011).

The results for measurement reports in relation to the students rated are shown in Table 3.

Table 3. Scored students' measurement report

	Measure	Standard error	Infit	Outfit
Mean	0.00	0.19	1.00	0.99
Standard deviation	0.68	0.08	0.33	0.46
Reliability = 0.91 Seperation = 3.16	chi-square = 1052.1	df = 84	p =0 .00	

On examining Table 3, it is clear that the average for students' performance scores is 0 and the standard deviation is 0.68 logit. Outfit statistics is the squares average of the residual between observed data and expected data, and it is sensitive to unexpected extreme values. For instance, it is an unexpected extreme value when a student receiving high scores from most of the performance tasks receives a low score from an easy performance task. Infit statistics is less sensitive to extreme values than outfit statistics, and it is desirable to have 1 as the infit value. Having an infit value above 1 indicates that a variance higher than the expected is observed whereas having the value below 1 indicates that a variance lower than the expected is observed (Güler, 2014; Hetherman, 2004). Having averages of 1 and 0.99 for infit and outfit values respectively indicates that model-data fit is almost perfect. On examining the discriminating rate and reliability shown in Table 3, it is found that the values are 3.16 and 0.91 respectively. Reliability value here receives a value between 0 and 1 as in classical test theory, and is interpreted in a similar way. Having a quite high value shows that students are discriminated adequately in terms of property measured.

The results for measurement reports in relation to the criteria used in rating in peer assessment are shown in Table 4.

Table 4. Criteria's measurement report

Criteria	Measure	Standard error	Infit	Outfit
C7	-0.55	0.04	0.71	0.78
C5	-0.83	0.04	0.93	0.93
C3	-0.94	0.04	1.54	1.31
C2	-1.38	0.05	1.23	1.13
C4	-1.49	0.06	0.63	0.64
C1	-1.54	0.06	1.32	1.15
C8	-1.55	0.06	0.90	1.08
C6	-2.57	0.10	1.48	0.94
C9	-3.14	0.14	1.10	0.90
Mean	-1.55	0.07	1.10	0.99
Standard deviation	0.78	0.03	0.31	0.19
Reliability = 0.99 Seperation = 11.24	chi-sqaure = 870.2	df = 8	p = 0.00	

On examining the measurement reports concerning the criteria in Table 4, it is clear that the most difficult criterion is C7 (variables) with the value of 0.55 logit and the easiest criterion is C9 (findings and interpretation) with the value of 3.14. The infit and outfit statistics receive values between 0.71 and 1.10, and between 0.78 and 0.90 respectively. The desired interval for fit statistics is between 0.5 and 0.90; and statistics outside this interval indicate weak fit (Wright & Linacre, 1994). Thus, it may be said that there are no criteria affecting the model-data fit in a negative way. Having a reliability value of 0.99 for the criteria- which is quite high- may be interpreted as that the criteria used in peer assessment functioned in a reliable way in discriminating students' performances.

The results for measurement reports in relation to the levels in the rubric (not appropriate "1", completely appropriate "5") used in peer assessment are shown in Table 5.

Table 5. Category statistics

Scoring levels	Counts used	%	Cumulative %	Average mean	Expected mean	Outfit
1	169	3	3	0.19	-0.01	1.3
2	180	3	7	0.33	0.31	1.2
3	475	9	15	0.58	0.68	0.9
4	995	19	34	1.10	1.16	0.8
5	3535	66	100	1.94	1.92	1.0

On examining the distribution of frequencies and percentages for the levels in the rubric (rating key) shown in Table 5, it is clear that lower levels are used rarely in assessment between 1 and 5, and that frequencies and percentages raise at higher levels. In order to be able to say that levels perform their function, distribution in rating levels should be balanced and there should be at least 10 students at each level (İlhan, 2016; Linacre, 2014). On examining the frequencies for the levels, it may be said that the desired values were reached. Besides, the fact that outfit values are close to 1 indicates that the levels perform their function effectively.

In Table 6, the students involved in peer assessment are divided into low, medium and high academic achievement levels; and the measurement report concerning the scores they give according to their achievement levels is also shown.

Table 6. Measurement report of grouped raters according to academic achievement level

Grouped raters	Measure	Standard error	Infit	Outfit
High	0.16	0.03	1.01	0.95
Medium	-0.03	0.03	1.17	1.16
Low	-0.13	0.03	0.94	0.84
Mean	0.00	0.03	1.04	0.98
Standard deviation	0.12	0.00	0.10	0.13
Reliability = 0.95	Seperation = 4.37	chi-square = 41.7	df = 2	p =0.00

According to Table 6, the logit value of the scores students with high academic achievement give to their peers is 0.16 (the strictest rating); and it is followed by -0.03 given by students with low achievement and -0.13 (the most generous rating) given by medium achievement. The reliability value for the scores given is 0.95 and discriminating rate is 4.37, and chi-square test for the difference between scores was found to be significant (Chi-square=41.7, $p<0.00$). Accordingly, it may be said that the scores given by students to their peers differ according to their academic achievement levels. In addition to that, significant differences were found according to one-way variance analysis in scores given by students according to their academic achievement ($F=6.576$, $p=0.002<0.00$), and consequently, post-hoc test was performed and thus it was found that the scores given by students with high achievement differed significantly from scores given by students with medium and low achievement.

Efforts were also made to identify students' tendencies in terms of peer assessment process through their answers to the open-ended questions in the form that they were asked to complete. In explaining the answers students gave to the questions, information on their achievement levels were also given at the end of the sentences in parentheses. In this way, correlations between students' views and their levels of achievement were described more clearly.

All of the students involved in peer assessment as raters said that they found the explanations made prior to the rating adequate and that the criteria included in the integrated rubric was clear and comprehensible. Almost all of the raters replied that they felt efficient when they asked whether or not they felt efficient in assessing their peers. One rater who felt partially efficient in rating said, "At the beginning I had doubts, but as I performed rating, I gained experience and I got used to it" (R3: a student with low level of achievement); and another rater said, "No, but I felt efficient after a few assessment trials" (R4: a student with medium level of achievement). As is clear from the example, students who had initially felt inefficient later felt efficient and began to display more positive tendencies in assessing their friends as they continued the assessment process.

When the raters asked to state their views of being objective in assessing their peers, they all said that they assessed their friends objectively.

When the raters were asked whether or not their achievement levels affected assessing their friends, all of them said that it did not affect their assessment. A student, for instance, said "my level of achievement did not affect my rating. I believed that I was fair in my treatment to my friends" (R7: a student with high level of achievement).

Raters were also asked whether or not they wished other students to assess them. That is to say, they were asked if they wanted to be the students assessed instead of being students who assessed others. Five of them said "yes". "Yes because my peers' assessing me gives them and me a different perspective" (R6: a student with medium level of achievement). "Yes. I would also like to see myself from the perspective of my peer" (R4: a student with medium level of achievement). One student, on the other hand, said that they would not like to be assessed by a peer saying "no because the teacher's assessment is a source of stress for me. When my friends also

assess me, it makes me tense" (R5: a student with high level of achievement).

Finally, raters' positive and negative views of peer assessment were obtained. As positive views, they stated that they understood their peers thanks to peer assessment, that peer assessment was not difficult, that it had effects on learning, and that it increased self-confidence. Quoted student views are as in the following:

- "Because we were peers, it was not difficult for us to understand what the other person felt and thought in the process of evaluation". (R7: a student with high level of achievement)
- "Making observations and learning through observations were effective". (R6: a student with medium level of achievement)
- "I have always found peer assessment stressful. When I became the rater, this situation made me see that it was not in fact negative". (R2: a student with low achievement level)
- "I became more informed of the topic of assessment". (R3: a student with low level of achievement)
- "It is good in that it strengthens students' interpersonal relations". (R4: a student with medium level of achievement)
- "Now I have increased self-confidence, I have new horizons about being fair." (R5: a student with high level of achievement)

As the negative sides of peer assessment, the students stated that it might cause them to compare themselves with their friends and that there were inadequacies in assessment. Quoted student views in this respect are as in the following:

- "There might be inadequacies in assessment because we receive the same education" (R7: a student with high level of achievement)
- "It can be disturbing in that students can compare themselves with others". (R4: a student with medium level of achievement)

4. Conclusion and Suggestions

This study analysed the reliability of scores given by six students and the lectures to 85 second year students attending the English Language Teaching department of a state university and taking the course of Scientific Research Techniques in the 2014-2015 academic year by means of MFRM. Besides, whether or not there were any differences between scores given by students participating in rating according to their levels of achievement (low-medium- high levels of achievement) was also researched in this study. The findings obtained through MFRM demonstrated that students' ratings and criteria for ratings were reliable. Besides, the fact that the infit and outfit statistics obtained in relation to the students and to the criteria were within the desired interval showed that model-data fit was available. On examining the results in the measurement reports concerning the levels (1-5) in the rubric, it was found that the outfit values were close to 1 and thus it was concluded that the levels were actively functional. Yet, the fact that distribution at lower levels (3%, 3% and 9% for 1, 2, 3 respectively) was few was considered as a reason to recommend that the levels could be arranged as few (for example in three levels) in similar studies. The facet of rater in the study was analysed in two different situations. Firstly, all raters (the lecturer and the peers) were included in the analysis, and thus the reliability of rating was analysed. In consequence, the strictest rating was performed by students with high level of achievement and the most generous rating was performed by students with low level of achievement. It was also found that the course lecturer was the third strictest rater. The fact that the fit statistics concerning the raters were within desired interval of values indicated that there were no raters damaging the model-data fit. The fact that the discriminating index between raters was 4.53 indicated that there were differences between raters. Similar studies available in the literature show that differences in the educational process raters go through and the differences in their levels of education, differences in their personality and personal history can lead to differences in rating (İlhan, 2016; Güler, 2014; Baştürk and Işıkoğlu, 2008; Semerci 2012; Brookhart, Walsh and Zientarski, 2006; Mulqueen, Baker and Dismuskes, 2000). In the other analysis performed for the fact of rater, the rating performed by the lecturer was not included, and the peers were divided into high, medium and low academic achievement level and were considered in these three categories. According to the result of the analyses done by means of MFRM, it was found that the students with high level of achievement performed the strictest rating while the students with medium level of achievement performed the most generous rating. As a result, it was found that the scores given by raters in peer assessment differed significantly according to their levels of academic achievement. After that, the post-hoc analysis was done and which groups had differences was analysed in this way. It was also found thus that the group with high level of academic achievement had score averages different from the groups with medium and low levels of academic achievement, and that there were no significant differences between scores of peers with low and medium level of achievement. In consequence, it may be said that MFRM could be effectively used in peer assessment especially in higher education.

And finally, the views held by students participating in peer assessment in relation to the process were obtained in this research. Although there were no seriously negative views in relation to the process, it was pointed

out that the process instilled in students various skills related to the topic of assessment and the practice of peer assessment.

Based on the findings obtained in this study it may be recommended that similar studies be conducted by using different measurement tools (observations, interviews, open-ended questions, surveys, etc.) and with raters having different properties.

References

Akın, Ö. & Baştürk, R. (2012). Keman eğitiminde temel becerilerin rasch ölçme modeli ile değerlendirilmesi. Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 31, 1, 175-187.

Aryadoust, V. (2015). Self-and Peer Assessments of Oral Presentations by Firs-Year University Students. Educational Assessment, 20: 199-225.

Baştürk, R. & Işıkoğlu, N. (2008). Okul öncesi eğitim kurumlarının işlevsel kalitelerinin çok-yüzeyli Rasch ölçme modeli ile analizi. Kuram ve Uygulamada Eğitim Bilimleri, 8(1), 7-32.

Bostock, S. (2009). Student Peer Assessment https://www.cs.auckland.ac.nz/courses/compsci747s2c/lectures/paul/Student_peer_assessment_-_Stephen_Bostock.pdf

Bozkurt, E., ve Demir, R. (2013). Öğrenci görüşleriyle akran değerlendirme: bir örnek uygulama. İlköğretim Online. 12(1), 241-253.

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

Büyüköztürk, Ş., Kılıç Çakmak, E. A., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2008). Bilimsel araştırma yöntemleri. Ankara: Pegem Akademi.

Çepni, S. (2007). Performansların değerlendirilmesi, E. Karip (Ed.) Ölçme ve Değerlendirme (193-239), Ankara: Pegem Yayıncılık.

Dochy, F. & McDowell, L. (1997) Assessment as a tool for learning, Studies in Educational Evaluation, 23, 279-298.

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

Falchikov, N. (1995) Peer feedback marking: developing peer assessment, Innovations in Education and Training International, 32, pp. 175-187.

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.

Freeman, M. (1995) Peer assessment by groups of group work. Assessment and Evaluation in Higher Education 20, 289-299.

Gibbs, G. (2006). Why assessment is changing? In K. Clegg & C. Bryan (Eds.), Innovative assessment in higher education (pp. 11–22): Routledge.

Güler, N. (2014). Analysis of Open-Ended Statistics Questions with Many Facet Rasch Model. Eurasian Journal of Educational Research. 55, 73-90.

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

Hetherman, S. C. (2004). An Application of Multi-Faceted Rasch Measurement to Monitor Effectiveness of the Written Composition in English in the New York City Department of Education. Unpublished Dissertation Thesis. Teacher College, Colombia University, Colombia.

İlhan, M. (2016). A Comparison of the Ability Estimations of Classical Test Theory and the Many Facet Rasch Model in Measurements with Open-ended Questions. Hacettepe University Journal of Education. DOI: 10.16986/HUJE.2016015182.

Karakaya, İ. (2015). Comparison of self, peer and instructor assessments in the portfolio assessment by using many facet Rasch model. Journal of Education and Human Development, 4(2), 182-192.

Kim, Y., Park, I. & Kang, M. (2012). Examining Rater Effects of the TGMD-2 on Children with Intellectual Disability. Adapted Physical Activity Quarterly, 29, 346-365.

Kutlu, Ö., Yıldırım, Ö. & Bilican, S. (2009). Öğretmenlerin dereceli puanlama anahtarına ilişkin tutum ölçeği geliştirme çalışması, Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, VI (II), 76-88.

Kutlu, Ö., Doğan, D. & Karakaya, İ. (2010). Öğrenci Başarısının Belirlenmesi, Ankara: Pegem Yayıncılık.

Landry, A., Jacobs, S. & Newton,G. (2015). Effective Use of Peer Assessment in a Graduate Level Writing Assignment: A Case Study. İnternational Journal of Higher Education. 4(1), 38-51.

Langan, A.M. & Wheater, P. (2003). Can students assess students effectively? Some insights into peer-assessment. Learning and Teaching in Action, 2(1).

Linacre, J. M. (2014). A user's guide to FACETS Rasch-model computer programs. Chicago, IL. [Available online at: http://www.winsteps.com/a/facets-manual.pdf], Retrieved on March 02, 2016.

Linacre, J. M. (2007). A User's Guide to FACETS: Rasch Model Computer Programs. Chicago, IL.
Linacre, J. M. (1989). Many-facet Rasch Measurement. Unpublished doctoral dissertation. University of Chicago, Chicago, IL.
Linacre, J. M. & Wright B. D. (2002). Construction of Measures from Many-facet Data. Journal of Applied Measurement, 3, 486-512.
Looney, M. A. (2012). Judging Anomalies at the 2010 Olympics in Men's Figure Skating. Measurement in Physical Education and Exercise Science, 16, 55-68.
Mulqueen C., Baker D. & Dismuske, R. K. (2000). Using Multifacet Rasch Analysis to Examine the Effectiveness of Rater Training. SIOP
Revesz, A. (2012). Working Memory and the Observed Effectiveness of Recasts on Different L2 Outcome Measures. Language Learning, 62, 1, 93-132.
Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. Language Testing, 25 (4), 553-581.
Sandvoll, R. (2014). When intentions meet reality: Consonance and dissonance in teacher approaches to peer assessment. Canadian Journal of Higher Education. 44(2), 188-134.
Semerci, Ç. (2012). Öğrencilerin BÖTE Bölümüne İlişkin Görüşlerinin Rasch Ölçme Modeline Göre Değerlendirilmesi (Fırat Üniversitesi Örneği). E-Journal of World Science Academy, NWSA-Education Sciences, 1CO542, 7(2), 777-784.
Semerci, Ç. (2011). Doktora yeterlikler çerçevesinde öğretim üyesi, akran ve öz değerlendirmelerin rasch ölçme modeliyle analizi. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 2(2), 164-171.
Stefani, A.J. (1994) Self, peer and group assessment procedures. In: An enterprising curriculum: Teaching innovations in Higher Education. Eds I. Sneddon and J. Kramer. Pp 24-46. HMSO, Belfast.
Topping, K. (1998) Peer assessment between students in colleges and universities. Review of Educational Research 68: 249-276.
Topping K. J. (2005). Trends in Peer Learning. Educational Psychology, 25(6), 631–645
Yurdabakan, H. (2011). Yapılandırmacı Kura. Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 44(1), 51-77.
Yurdabakan, H. ve Cihanoğlu, M., O. (2009). Öz ve akran değerlendirmenin işbirlikli okuma ve kompozisyon tekniğinin başarı, tutum ve strateji kullanımına etkisi. Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 11(4), 105-123.
Yurdabakan, İ., ve Oğlun, M. (2011). Öz ve akran değerlendirmenin öğrenme ve bilişüstü bilgi üzerindeki etkisi: sonuçsal geçerlik. 2nd International Conference on New Trends in Education and Their Implications 27-29 April, 2011 Antalya-Turkey.Lawrence, S. et al. (2001). Persistence of Web References in Scientific Research. Computer. 34, 26-31. doi:10.1109/2.901164, http://dx.doi.org/10.1109/2.901164
Smith, Joe, (1999), One of Volvo's core values. [Online] Available: http://www.volvo.com/environment/index.htm (July 7, 1999)
Strunk, W., Jr., & White, E. B. (1979). The elements of style. (3rd ed.). New York: Macmillan, (Chapter 4).
Van der Geer, J., Hanraads, J. A. J., & Lupton R. A. (2000). The art of writing a scientific article. Journal of Scientific Communications, 163, 51-59