An analysis of scoring via analytic rubric and general impression in peer assessment

Nagihan Boztunç Öztürk [iD]
Hacettepe University, Lifelong Learning Center, Ankara,, Turkey, nagihanboztunc@hacettepe.edu.tr
Melek Gülşah Şahin [iD]
Gazi University, Gazi Education Faculty, Ankara, Turkey, mgulsahsahin@gazi.edu.tr
Mustafa İlhan [iD]
Dicle University, Ziya Gökalp Education Faculty, Diyarbakır, Turkey, mustafa.ilhan@dicle.edu.tr

ABSTRACT   The aim of this research was to analyze and compare analytic rubric and general impression scoring in peer assessment. A total of 66 university students participated in the study and six of them were chosen as peer raters on a voluntary basis. In the research, students were supposed to prepare a sample study within the scope of scientific research methods course and were also expected to present their studies in class. While the students were giving a presentation, their course instructor and peer raters conducted scoring, firstly by using the analytic rubric and subsequently by using general impressions. Collected data were analyzed using the Rasch model. Consequently, it was found that students were distinguished from one another at a highly reliable rate using both scoring methods. Additionally, it was discovered that the differences between students' ability levels were better revealed when analytic rubric was used. It was ascertained that there was a high level positive correlation between the ability estimations obtained from the scores performed by the peers and the instructor, regardless of the scoring method used. Finally, it was determined that ability estimations, corresponding peer raters' analytic rubric and general impression scoring, held a positive and highly strong relation.

Keywords: *Peer assessment, Analytic rubric scoring, General impression scoring, Many-facet Rasch model.*

Akran değerlendirmede analitik rubrikle ve genel izlenimle yapılan puanlamaların incelenmesi

ÖZ   Bu araştırmada, akran değerlendirmede analitik rubrikle ve genel izlenimle yapılan puanlamaların karşılaştırmalı olarak incelenmesi amaçlanmıştır. Araştırma 66 üniversite öğrencisi üzerinde yürütülmüş ve bu öğrencilerden gönüllülük esasına dayalı olarak seçilen altısı, çalışmada akran değerlendirici olarak görev almıştır. Çalışmada, öğrencilerden bilimsel araştırma yöntemleri dersi kapsamında örnek bir çalışma hazırlamaları ve hazırladıkları çalışmayı sınıf ortamında sunmaları istenmiştir. Öğrenciler sunum yaparken dersin sorumlu öğretim elemanı ile akran değerlendiriciler, çalışmaları önce analitik rubriğe göre ve ardından genel izlenimle puanlamıştır. Puanlamadan elde edilen veriler, Rasch modeline göre analiz edilmiştir. Araştırmada, her iki puanlama yönteminde de bireylerin yüksek güvenirlik birbirinden ayırt edildiği belirlenmiştir. Bununla birlikte, analitik rubrik kullanıldığında bireylerin yetenek düzeyleri arasındaki farklılıkların daha hassas bir biçimde ortaya konulduğu saptanmıştır. Hem analitik rubrikle hem de genel izlenimle yapılan değerlendirmede; akranlar ile öğretim elemanının verdiği puanlar üzerinden hesaplanan yetenek kestirimleri arasında, pozitif yönlü yüksek korelasyonlar bulunmuştur. Akranların analitik rubrikle ve genel izlenimle yaptıkları puanlamalara karşılık gelen yetenek kestirimlerinin pozitif yönlü güçlü bir ilişki içerisinde olduğu sonucuna varılmıştır.

Anahtar   *Akran değerlendirme, Analitik rubrikle puanlama, Genel izlenimle puanlama, Çok yüzeyli Rasch*
Kelimeler: *modeli.*

258

INTRODUCTION

Learning is an ongoing activity that takes place in everyday life independently from time and space. People might require and seek guidance from their friends and colleagues about specific issues in their working or private lives, and thus might exchange ideas. Such a process —despite not being the most effective way of learning, and despite it not resulting in true knowledge acquisition all the time— nevertheless offers some important advantages. Of these advantages, the most leading is that which is included in Bandura's social learning theory, and which is the contribution of the similarity between the learner's (observer) and the teacher's (model) positions to learning.

Learning from one another not only occur only at home, in the street, or at work, it also takes place within formal education settings. When students get stuck when studying a problem, they often initially try to find a solution by resorting to classmates or other friends instead of consulting their teacher. In this sense, formalized peer education has great potential in terms of enhancing learning effectiveness and, with the help of peer education, students assume the responsibility of their own learning, have the opportunity to study with others, and are given the chance to engage more practical learning when compared with traditional teaching-learning methods. Consequently, peer education helps students become skillful at critical questioning, reflection, communication, managing learning, as well as in self- and peer-assessment (Boud, 2013). Those qualities mentioned above accord with 21st century skills indicating, to a great extent, the importance of peer education as part of the teaching-learning processes.

Peer Assessment

One of the concepts that falls under peer education is peer assessment. Various definitions of peer assessment have been devised by different researchers. According to McDonald (2016), in simplistic terms, peer assessment is an assessment by peers or colleagues with the purpose of identifying the quality of work completed. Topping (1998) defines peer assessment as an arrangement in which students consider the amount, level, value, worth, quality, or success of a product or the performance of other similar-status learners. Comparatively, Falchikov (2001) conceptualizes peer assessment students' evaluation of their peers' products or performance in line with previously designated criteria. As is clear from these definitions, within peer assessment, students evaluate their friends' works on the one hand, while obtaining feedback from their friends and comparing their works with those of their friends on the other.

Peer assessment is used often in higher education, as is the case across other levels of education (Amo & Jareno, 2011; van den Berg, Admiraal, & Pilot, 2006). Studies carried out in the academic field set forth the common usage of peer assessment in higher education (Hanrahan & Isaacs, 2001; Macpherson, 1999; Mehrdad, Bigdelib, & Ebrahimia, 2012; Smith, Cooper, & Lancaster, 2002; Şahin, Taşdelen-Teker, & Güler, 2016; Taşdelen-Teker, Şahin, & Baytemir, 2016; Wen & Tsai, 2006). Sluijmans, Brand-Gruwel, van Merrienboer, and Bastiaens (2003) underline the fact that the use of self-assessment, as well as peer assessment, is gradually increasing in teacher-training institutions as both methods comply with contemporary pedagogical approaches related to training teacher candidates.

Numerous advantages are facilitated by having peers assess each other's work. First, peer assessment creates a learning culture that is based on increased participation and collaborative understanding (Mutwarasibo, 2016); it generates a learning environment in which students can assume greater responsibility for their own learning and thus facilitates students' learning autonomy (Ashraf & Mahdinezhad, 2015). It also supports development of students' communication (Gravells, 2014), critical-thinking (Tan, 2015), and decision-making skills, as well as other meta-cognitive skills (Berry,

259

2008). Furthermore, it presents an effective feedback for students (Topping, 2009), and encourages deep learning instead of superficial learning (Karaca, 2009). Peer assessment popularizes the idea of assessment for learning and, in this way, mistakes are accepted as opportunities for improving learning further, not as failures (Bostock, 2000). In addition to all these advantages, Prins, Sluijsmans, Kirschner, and Strijbos (2005) state that those skills that students are expected to acquire through peer assessment are necessary in various professional contexts, and so they claim that peer assessment will help students gain lifelong learning skills and better prepare them for working life.

Undoubtedly, peer assessment brings about certain restrictions in addition to the various advantages specified above. The limitations of peer assessment include the argument that it might be challenging to create an appropriate environment and the necessary conditions; that some students who are expected to score do not have enough self-confidence regarding assessment; that peer assessment can lead to anxiety and stress for some students; that, within peer assessment, it is difficult to ensure that all students who are to undertake scoring of their peers have an equivalent understanding of the rating criteria; that peer assessment need to be supported by other assessment methods (Gravells, 2014); and that assessments and feedback of peers might be underestimated by the individual whose work is being assessed (Yakar, 2019). However, beyond all the aforementioned points, the most controversial point regarding peer assessment concerns the problems of validity and reliability (McDonald, 2016). The validity and reliability of peer assessment are negatively influenced by the fact that the students who conduct the assessment may not possess the necessary content knowledge and skills (Mutwarasibo, 2016), that students may not act objectively, and that so in-class friendship relations might be reflected in scoring (Mann, 2006).

In those studies, which aim to designate the reliability and validity of peer assessment, grades given by peers are often compared and judged according to the tutor's marks (Frankland, 2007). If the tutor's assessment is assumed to be reliable and valid, then a high level of agreement between the scores given by peers and those given by the tutor indicates that peer assessment is accurate (Topping, 2009). On reviewing previous studies, Topping, Smith, Swanson and Elliot (2000) discovered that there was a high level of similarity between positive and negative statements in assessments undertaken by peers and instructors. Falchikov and Goldfinch (2000) carried out a meta-analysis using 48 studies that compared peer and teacher scoring, and found out that students made judgements that were reliable at a reasonable level. In another study, Şasmaz Ören (2018) examined the relationships between self, peer, and teacher assessments; it was determined that there was a moderately high correlation between peer- and teacher-assessment scores. Matsuno's (2009) study indicated that most peer-raters were internally consistent and produced fewer bias interactions than teacher-raters. Additionally, Topping (2009) stated that peer assessment produces more reliable results when supported by training, checklists, exemplification, teacher assistance, and monitoring.

Peer Assessment for Formative and Summative Purposes

Peer assessment can be used for formative- (monitoring progress during instruction) and/or summative-assessment (assigning grade) purposes. In summative assessment, peers are expected to give a mark or help in grading; comparatively, in formative assessment, peers are expected to provide feedback/comments. At this point, it is worth noting that there is no obligation to choose either one of these two approaches. It is also possible to use two approaches in combination if peers are asked to give both grades and feedback (Liu & Carless, 2006). However, studies have shown that formative purpose peer assessment tends to give more accurate results compared with summative-purpose peer assessment (O'Donnell & Topping, 1998 cited in McLeod, Brown, McDaniels, & Sledge, 2009).

Formative and summative assessment differ in terms of scoring methods used as well as in regard to their purposes. Summative assessment includes different grading methods, one of which is scoring using general impression. In general impression scoring, the rater reads the paper in its entirety before giving a single score in consideration of the grading system used. In this scoring type, rater works with no written criteria and no detailed explanation is provided regarding the given score (Lester, Lambdin, &

260

Preston, 1997). Comparatively, in formative assessment, scoring keys such as checklists, rating scales, and rubrics are used. Rubrics are the most commonly used of these scoring keys.

According to Kan (2007), the rubric is a scoring guide that defines the characteristics and criteria of different levels of performance, and is used to make judgments about performance in accordance with certain characteristics and criteria. That is, rubrics are typically employed when a judgement of quality is required and may be used to evaluate a broad range of subjects and activities (Moskal, 2000). Rubrics provides more reliable scoring on the one hand (Moskal & Leydens, 2000) and serve as a means of communication between teachers, students, and parents about the strengths and weaknesses of students on the other (Hall & Salmon, 2003). Rubrics are divided into holistic and analytic rubrics in accordance with the scoring strategy they employ. A holistic rubric is based on a global impression of the performance or product. The measured structure is not subdivided when scoring using a holistic rubric (Gronlund, 1998). More clearly, the student's performance or product is evaluated as a whole and given a single score (Nitko, 2004). In other respects, in the analytic rubric performance is divided into components and each component of the performance is scored separately (Reddy, 2010). Subsequently, the scores given for each component are summed to obtain an overall performance score (Petkov & Petkova, 2006).

Aim and Significance of the Study

In today's world, it is necessary to make use of assessment systems that help to develop individuals who have lifelong learning skills, and who are equipped with skills prescribed by the era of information and technology in higher education, as well as in other levels of education. It is especially important to include such assessment methods in faculties of education, as these are the institutions responsible for educating future teachers. Within such a context, it can be plainly stated that there is a need to resort to assessment approaches that will help individuals to acquire 21st century skills in higher education in general and, more specifically, within faculties of education. One of the leading approaches that has the potential of meeting the above-mentioned needs comprises student assessment methods within the measurement and evaluation process, one of which is peer assessment. Therefore, it is of great importance to carry out studies that address peer assessment in higher education according to various aspects, that reveal the conditions that are to be used, and that will provide more functional and accurate results. Within this framework, the aim of this study is to examine scoring via analytic rubric and general impression in peer assessment using a sample of higher education students. In this direction, it is intended that this study will: *i)* designate reliability values regarding peer scorings through analytic rubric and general impressions; *ii)* analyze the correlation coefficients between ability estimations calculated according to the peer raters' and instructor's scores separately for the two scoring methods *iii)* test the agreement between the ability estimations corresponding peer raters' analytic rubric and general impression scoring. On examination of the literature, it can be seen that studies aimed at comparing different scoring methods are mostly conducted on teachers/expert raters. No comparative study on the investigation of scoring validity and reliability using an analytic rubric and general impression in peer assessment was found. This highlights the original value of this research and that it will make a significant contribution to the literature.

METHODOLOGY

Participants

The participants of this study comprised 66 third-year university students at the Department of Early Childhood Education at a state university in Turkey during the 2018-2019 academic year. The study was conducted in one of the classes where the researcher lectured. Hence, it can be said that the

261

participants were determined according to the convenience sampling technique. In this research, peer assessments were made through the studies prepared by the students within the scope of scientific research methods course. Scientific research methods course is a compulsory course taught two hours a week for one semester. In the period of the study, 66 students who have to attend this course consisted the participants of the study. Students who failed in the course in the previous semesters and only had to take the exams of the course were not included in the study. Students were supposed to prepare a sample research within the scope of this course and were also expected to present their research in class. Six of the students (four female and two male) were chosen on a voluntary basis to carry out the peer assessment. These six students were named as peer raters in the following sections of the study. In deciding the number of peer raters, the results of Falchikov and Goldfinch's (2000) meta-analysis research were taken as the basis. Falchikov and Goldfinch (2000) reported that the correlations between peer raters were higher when the number of peers was between 2 and 7. Also, attention was paid to the fact that the peer raters are of different achievement levels and two students from each of the low, medium and high achievement levels were selected based on their overall grade point averages.

Data Collection Tools

The data collection tool used in this study was an analytic rubric comprising 13 dimensions and had been developed for assessing the quality of scientific studies prepared by students. This rubric, which adopted a three-grade (0-1-2) approach, was developed by the researchers for this study. After a draft form of the rubric had been prepared, the form was then presented to three experts who were academicians in the field of measurement and evaluation. One of the experts stated that, the research purpose and the problem sentence of the research was expressed in a single dimension in the analytic rubric and that these two elements must be arranged in two different dimensions. Considering this recommendation of the expert the two elements just mentioned are arranged in two separate dimensions. Another expert stated that the phrase of "*necessary rules*" in the sentence of "*The sub-problems were expressed in accordance with the necessary rules*" was not clear. That's why, the sentence was changed as "*The sub-problems are clear, understandable and consistent with the research purpose*".

After the necessary amendments and changes had been implemented in line with the suggestions made by these experts; the views of two more experts in the field of measurement and evaluation were consulted, who deemed the rubric was ready for use. That the rubric had three grades was also supported, not only by the views of the experts, but also by the category statistics reported in Rasch analysis. The category statistics obtained by analyzing the data of analytic rubric scoring are given in Table 1.

Table 1.
Results of category statistics regarding the three-grade in the analytic rubric

Category Score	Counts	Percent	Cumulative Percent	Average Measure	Expected Measure	Outfit MnSq
0 (Inadequate)	602	13	13	-.09	-.04	.90
1 (Acceptable)	1117	24	37	.83	.78	1.10
2 (Good)	2961	63	100	1.70	1.71	1.00

There are several conditions that needed to be met in order to determine whether those categories used in the rubric are appropriate, and that they could be distinguished by raters without issue: *i)* there should be at least 10 observations in each rubric category, and observation distribution should be regular, *ii)* the average measures should advance monotonically with rubric categories; and *iii)* outfit mean-squares should be less than 2.0 (Linacre, 2002). As can be seen in Table 1, these three conditions are met in the current study. Therefore, it can be stated that three-level rubric used in this study worked without issue.

Procedure

In this study which is an applied research, the data was obtained from the applications of one of the researchers in the scientific research methods course. Peer assessment is an assessment method that the researcher has already included in the scope of the course, and apart from that, no measurement tool was

262

applied to the students. In the study, the steps followed in the peer assessment process can be summarized as follows: Researchers came together with the six chosen students and provided them with training prior to scoring. During this training, peer raters were instructed about the qualities and necessities of peer assessment. Moreover, the analytic rubric was introduced to the peer raters, and a scoring session undertaken using a sample research with necessary explanations. Subsequently, students were told how to conduct the scoring via general impressions. Particular emphasis was given to the vital importance of not being affected by analytic rubric scores when scoring via general impressions; students were also reminded to close the analytic rubric scoring while scoring via general impressions.

Fundamentally, it can be seen as a more accurate way of scoring with general impression first to ensure that the analytic rubric scoring does not affect the general impression. However, following such a path requires each student to present his/her study a second time in order to be able to perform analytic scoring following the general impression. When considered from this point of view, doing the general impression scoring firstly is not useful. Therefore, performing analytic rubric scoring firstly is more applicable economically.

After deciding the path to follow in scoring, both peer raters and those students undertaking the presentations were informed that the data gathered would be used only within the scope of this research. Peer raters were told that they did not have to write their names on the assessment sheets, and that they could use a nickname while scoring if they wanted, and students who made the presentations were asked to undertake their presentations according to the titles determined by the researchers considering the dimensions in the analytic rubric. While the students were giving a presentation, their course instructor and peer raters conducted scoring, firstly by using the analytic rubric and subsequently by using general impressions for each and every student.

Data Analysis

Research data were analyzed according to the Rasch model. Before the results regarding the Rasch analysis were interpreted, the researchers checked to see whether analysis assumptions had been met. Rasch analysis entail three assumptions: unidimensionality, local independence, and model-data fit. Although these assumptions are expressed under separate titles, they are not independent from one another. Local independence functions in parallel with unidimensionality (Hambleton, Swaminathan, & Rogers, 1991), while unidimensionality is justified with the model-data fit (Lee, Peterson, & Dixon, 2010). To clarify, model-data fit indicates that the assumption of unidimensionality is ensured, while ensuring the assumption of unidimensionality indicates that there is no problem about the assumption of local independence. In this sense, analyzing if the model-data fit is present or not is the basic assumption that must be tested within Rasch analysis (Güler, İlhan, Güneyli, & Demir, 2017).

In Rasch analysis, standardized residuals are used in order to test the model-data fit. According to Linacre (2018), in order to say that there is a model-data fit, standardized residuals out of the ±2 interval should not exceed approximately 5% of the total data number, while standardized residuals out of the ±3 interval should not exceed approximately 1% of the total data number. However, McNamara (1996) states that when, analyzing the fit between model and data, those criteria suggested by Linacre (2018) do not have to be obeyed so strictly. According to McNamara (1996), Rasch model should not be abandoned for performance assessment as long as the percentage of the standard residuals left out of ±2 or ±3 interval does not remarkably deviate from the suggested criteria.

Two separate Rasch analyses were carried out in this study. The first analysis was undertaken using a data set concerning the scoring via analytic rubric. In this data set, which comprises three facets, the number of data was 468 (6x60x13), as there were six peer raters, 60 students, and 13 dimensions within the analytic rubric. On examination of the standardized residuals, the number of data that fell out of the ±2 interval was found to be 246 (5.26%), while the number of data that fell out of the ±3 interval was found to be 77 (1.65%). Even though these values did not exactly correspond those criteria suggested

263

by Linacre (2018), they did not deviate from the related values to a great extent. Therefore, it is possible to say that there was an acceptable fit between model and data.

The second Rasch analysis was undertaken using the data set regarding general impression scoring. In the general impression scoring made by the instructor, the only variability source is the students. Therefore, Rasch analysis could not be performed in the general impression scoring of the instructor and the instructor's grades were taken as students' ability measures. On the other hand, there are two variability sources in the general impression scoring of peer raters: students and peer raters. Accordingly, Rasch analysis with two facets was performed on this data set. Scoring was not done on the basis of dimension and one score was given according to general impressions about students' performance; accordingly, the facet of dimension was not included in the analysis. As there were six peer raters and 60 students in the data set, the total number of data was 360 (6×60). When standardized residuals were examined, no value was found that fell out of the ±3 interval, whereas 15 (4.16%) standardized residuals fell out of the ±2 interval. Accordingly, it can be understood that the model-data fit was justified.

For both sets of data, the fit between the model and the data indicates that the assumption of unidimensionality has been met, and that this consequently proves that the assumption of local independence is ensured. After it was decided that these assumptions had been met, reliability coefficients, separation indexes, and chi-squared values regarding the student and peer rater facets of analytic rubric and general impression scoring were comparatively examined, in line with the first sub-problem of this study. Within the scope of the second sub-problem, the correlation coefficients (Pearson product-moment correlation coefficient) between ability estimations obtained from peer raters' and the instructor's scoring, were calculated separately for the two scoring methods. Likewise, correlation analysis was carried out for the third sub-problem of this study in order to test the agreement between ability estimations that correspond to peer raters' scorings performed with analytic rubric and general impression. In the study, the FACETS 3.70.1 software was used for Rasch analysis, whereas correlation analysis was done using the SPSS 21.0 software program.

RESULTS

In presenting the results obtained in the study, the variable map reported at the end of the Rasch analysis was given first. Figure 1 shows the variable map concerning the many-facet Rasch analysis based on the analytic rubric scores.

264

```
+--------------------------------------------------------------------+
|Measr|+STUDENT                |-DIMENSION          |-PEER RATER|Scale|
|-----+------------------------+--------------------+-----------+-----|
|  4  +                        +                    +           +  (2)|
|     |                        |                    |           |     |
|     |                        |                    |           |     |
|     |  52                    |                    |           |     |
|     |                        |                    |           |     |
|  3  +  53  58                +                    +           +     |
|     |                        |                    |           |     |
|     |                        |                    |           |     |
|     |  34  57                |                    |           |     |
|     |  18  20  43  46  54  55|                    |           |     |
|  2  +  7   19  21  30        +                    +           +     |
|     |  25  35  36  44  51    |                    |           |     |
|     |  1   6   11  23  24  26  29  60 |           |           |     |
|     |  3   5   10  40  42    | 2                  |           |     |
|     |  2   14  37  45        |                    |           |     |
|  1  +  48                    +                    +           + --- |
|     |  27  41                |                    |           |     |
|     |  4   8   16  22  38  49  50 | 10            | 2         |     |
|     |  59                    |                    | 3    4    |     |
|     |  12  15  17  56        | 6                  |           |     |
*  0  *  9   13  31  32  39    * 11  3   5   7   9  *           *  1  *
|     |  28  47                | 12  4              | 5         |     |
|     |  33                    | 13                 |           |     |
|     |                        | 1                  | 1    6    |     |
|     |                        |                    |           |     |
| -1  +                        + 8                  +           + --- |
|     |                        |                    |           |     |
|     |                        |                    |           |     |
|     |                        |                    |           |     |
|     |                        |                    |           |     |
| -2  +                        +                    +           +  (0)|
|-----+------------------------+--------------------+-----------+-----|
|Measr|+STUDENT                |-DIMENSION          |-PEER RATER|Scale|
+--------------------------------------------------------------------+
```

Figure1. Variable map resulting from the many-facet Rasch analysis for analytic rubric scores

The 60 students in the study group were listed according to their ability levels, as can be seen in the second column of Figure 1. The fact that the students show a wide range of distribution in the column indicates that the students with different ability levels were effectively distinguished. There were 13 criteria of the analytic rubric in the dimension column of the variable map. The fact that these dimensions did not heap together at one single point, and that they were located at different points of the variable map, reflects that the 13 dimensions in the rubric differ in terms of their difficulty levels, and that the peer raters were able score the students' performances in different dimensions of the measured structure independently from one another. As can be seen in the peer-rater column (Figure 1), six peers who scored students' performances were not located at one point in the variable map, although they did not show a wide distribution range. Accordingly, it is obvious that peer raters differed from one another in terms of their severity and leniency. After the variable map regarding the scoring via analytic rubric had been looked into, the variable map concerning the general impression scoring was then examined (see Figure 2).
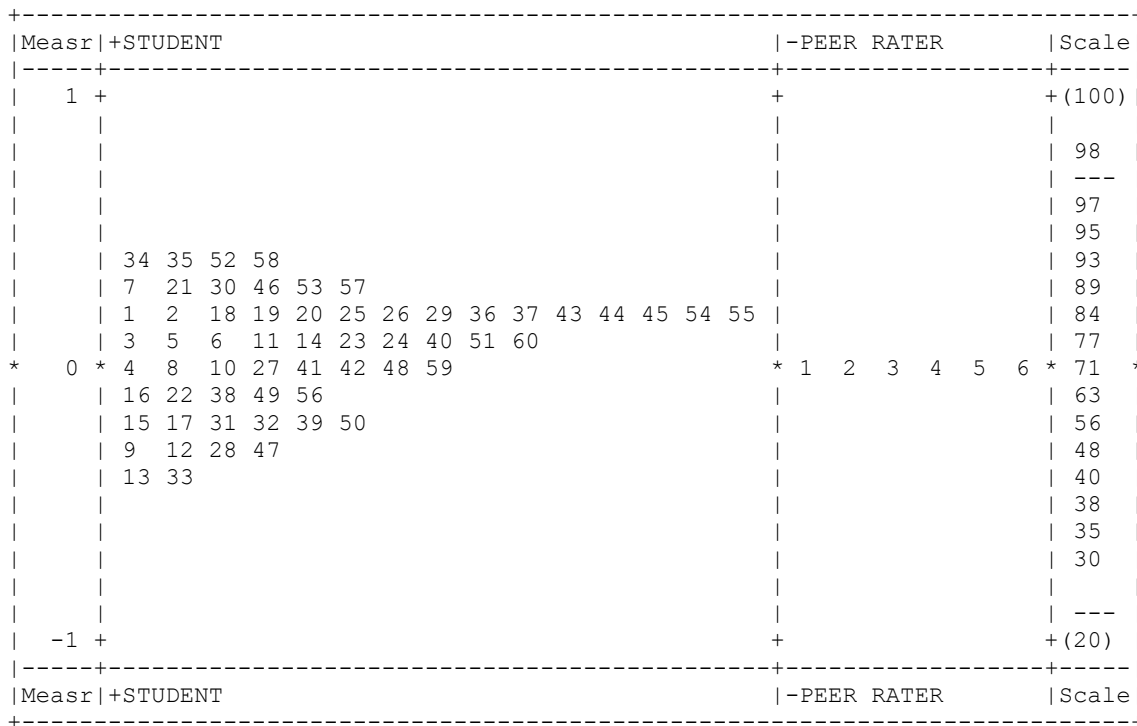
```
+-------------------------------------------------------------------------+
|Measr|+STUDENT                                        |-PEER RATER    |Scale|
|-----+------------------------------------------------+---------------+-----|
|  1  +                                                +               +(100)|
|     |                                                |               |     |
|     |                                                |               | 98  |
|     |                                                |               | --- |
|     |                                                |               | 97  |
|     |                                                |               | 95  |
|     | 34 35 52 58                                    |               | 93  |
|     | 7  21 30 46 53 57                              |               | 89  |
|     | 1  2  18 19 20 25 26 29 36 37 43 44 45 54 55   |               | 84  |
|     | 3  5  6  11 14 23 24 40 51 60                  |               | 77  |
*  0  * 4  8  10 27 41 42 48 59                        * 1  2  3  4  5  6 * 71  *
|     | 16 22 38 49 56                                 |               | 63  |
|     | 15 17 31 32 39 50                              |               | 56  |
|     | 9  12 28 47                                    |               | 48  |
|     | 13 33                                          |               | 40  |
|     |                                                |               | 38  |
|     |                                                |               | 35  |
|     |                                                |               | 30  |
|     |                                                |               |     |
|     |                                                |               | --- |
| -1  +                                                +               +(20) |
|-----+------------------------------------------------+---------------+-----|
|Measr|+STUDENT                                        |-PEER RATER    |Scale|
+-------------------------------------------------------------------------+
```

Figure 2. Variable map resulting from the Rasch analysis for general impression scores

As can be seen on examination of Figure 2, it is clear that unlike the variable map in Figure 1, there is no dimension column. The reason for this is that, during general impression scoring, the Rasch analysis was carried out according to two facets—students and peer raters—while the facet of dimension was not included in the analysis. When the column of students, as seen in Figure 2, is compared to the one in the variable map belonging to the analytic rubric scoring, it is obvious that students gathered in narrower intervals in terms of their ability levels. In other words, the range value concerning students' ability levels is lower for general impression scoring as compared with that of analytic rubric scoring. Consequently, it can be said that students with different ability levels were able to distinguish from one another better when using the analytic rubric. As can be seen in the peer-rater column, the six peers who conducted scoring were located at the 0 level of the variable map. According to their position in the variable map, no significant difference was found among peer raters in terms of their severity and leniency in general impression scoring. These inferences obtained from the variable maps, via analytic rubric and general impression scoring are also supported by the measurement reports given in Table 2.

Table 2.
Measurement reports reached at the end of the Rasch analysis for analytic rubric and general impression scorings

	Student Facet		Peer Rater Facet		Dimension Facet	
	Analytic rubric	General Impression	Analytic rubric	General Impression	Analytic rubric	General Impression
Infit MnSq	1.02	.99	.99	1.00	.98	
Outfit MnSq	1.03	.99	1.03	.99	1.03	
Separation	3.88	3.27	8.37	.25	6.01	-
Reliability	.94	.91	.99	.06	.97	
df	59	59	5	5	12	
Chi-square	992.70**	581.10**	348.80**	5.30*	476.40**	

$^{**}p < .05, ^{*}p > .05$

According to the results in Table 2, fit statistics in scorings via both analytic rubric and general impression fall within the acceptable interval of .5 and 1.5 (Wright & Linacre, 1994) for all the facets included in the analysis. The fact that fit statistics fall within the suggested interval indicates that model-data fit was ensured, and proves the validity of the measures. As is seen in Table 2, separation index,

266

reliability coefficient, and chi-square values belonging to the student facet are higher for analytic rubric scoring as compared with general impression scoring. Considering these values, it can be said that the differences between students' ability levels are revealed better when scoring is performed using the analytic rubric. However, it should also be noted that students with different ability levels can be distinguished from one another with high reliability using general impression scoring. On close examination of the peer rater facet, it is obvious that there is a statistically significant difference between raters in terms of severity and leniency when using analytic rubric scoring; no such difference is found regarding general impression scoring. Concerning general impression scoring, raters do not carry out an assessment on the basis of dimensions, rather, they only provide a single score based on their general impression about the student's performance. Therefore, analysis results regarding the dimension facet are available only for analytic rubric scoring. The fact that chi-square value calculated for the dimension facet is significant, and that separation index and reliability coefficient are high, indicates that the analytic rubric comprises criteria that have various levels of difficulty, and that students' performances across different dimensions of the measured structure were distinguished by peer raters. Following the measurement reports, correlations between the ability estimations, calculated for the scores given by the peers and the instructor, were examined. The correlation analysis results are given in Table 3.

Table 3.
Correlation coefficients between the ability estimations calculated according to peer raters' and the course instructor's scorings

Scoring method	Analytic rubric	General Impression
r	.718[**]	.723[**]

[**]*p* < .05

Hinkle, Wiersma, and Jurs (1979) state that, according to absolute value, if the correlation coefficient is between .00 and .30, it is very low; if it is between .30 and .50, it is low; if it is between .50 and .70, it is at medium level; if it is between .70 and .90, it is strong and if it is over .90, it is very strong. When the values in Table 3 are examined in respect to these intervals, it is obvious that there is a strong relationship between the ability estimations obtained from the scores performed by the peers and those obtained from the instructor, regardless of whether analytic rubric or general impression scoring methods.

The last finding of the study concerns the agreement between the ability estimations that correspond to analytic rubric and general impression scores given by the peer raters. While grading between 0 and 2 was used in the analytic rubric, scoring according to the general impression ranged between 0-100. This difference also affected the correspondent values of the results regarding the two types of scoring in the logit unit. Therefore, it was thought that it would not be appropriate to make an absolute comparison between ability estimations using scores gathered via the analytic rubric and general impression. Consequently, analyzing the agreement between the ability estimations calculated via the two types of scoring was restricted to relative agreement. Moreover, since the focus of the study was peer assessment, it was not deemed necessary to compare the ability estimations based on the instructor's analytic rubric and general impression scorings. The ability estimations calculated in scoring with the analytic rubric and general impression and, correlation analysis result to determine the relative agreement between these ability estimations are given in Table 4.

267

Table 4.
Ability estimations calculated using the scores via analytic rubric and general impression methods, and the correlation between these ability estimations

Student Number	Ability Estimations		Student Number	Ability Estimations		Student Number	Ability Estimations	
	Analytic rubric	General Impression		Analytic rubric	General Impression		Analytic rubric	General Impression
S1	1.66	.23	S21	1.92	.28	S41	.84	.02
S2	1.20	.16	S22	.61	-.11	S42	1.39	-.02
S3	1.32	.12	S23	1.52	.09	S43	2.25	.25
S4	.53	.00	S24	1.66	.15	S44	1.86	.23
S5	1.35	.11	S25	1.75	.22	S45	1.20	.15
S6	1.66	.12	S26	1.66	.21	S46	2.10	.25
S7	2.04	.27	S27	.84	.04	S47	-.23	-.26
S8	.59	.01	S28	-.13	-.28	S48	.94	.01
S9	.07	-.26	S29	1.66	.22	S49	.53	-.09
S10	1.32	-.01	S30	1.92	.32	S50	.64	-.23
S11	1.56	.10	S31	.09	-.22	S51	1.86	.11
S12	.14	-.26	S32	.07	-.18	S52	3.3	.40
S13	-.08	-.39	S33	-.47	-.37	S53	2.96	.26
S14	1.17	.07	S34	2.41	.36	S54	2.10	.25
S15	.30	-.19	S35	1.81	.36	S55	2.10	.20
S16	.51	-.11	S36	1.75	.23	S56	.12	-.10
S17	.12	-.22	S37	1.28	.16	S57	2.50	.25
S18	2.10	.22	S38	.67	-.08	S58	2.96	.35
S19	2.04	.25	S39	-.03	-.22	S59	.43	-.02
S20	2.25	.21	S40	1.32	.09	S60	1.61	.08

n = 60, *r* = .93, *p* < .05

According to the correlation coefficient, as seen in Table 4, there is a positive and strong statistically significant relationship between calculated ability estimations when the same performance is scored with analytical rubrics and general impression (*r* = .93; *p* < .05).

DISCUSSION, CONCLUSION and SUGGESTIONS

The aim of this study was to analyze and compare peer assessment scoring undertaken using an analytic rubric and general impression methods. First, results reported in the Rasch analysis concerning the student- and peer-rater facets were comparatively examined for both scoring methods, in accordance with the framework of this study. This comparison revealed that students were distinguished from one another at a highly reliable rate using both scoring methods. However, it was found out that the differences between students' ability levels were better revealed when using the analytic rubric. Accordingly, it would be better choice to use the analytic rubric—instead of general impression—method when scoring if there is an assessment in which the small differences between students' ability levels have the potential to change the decisions to be taken. Comparatively, scoring via on general impression may be a more economical scoring choice if small differences do not change the students' ranking in regard to their level of ability. The research findings for the student facet are in parallel with the results of the studies carried out by Alharby (2006) and Wiseman (2008, 2012). In their researches, Alharby (2006) and Wiseman (2008, 2012) carried out Rasch analysis, making use of scores obtained via analytical and holistic rubric and finding out that chi-square, separation, and reliability values calculated for the student facet were higher when analytic rubric was used instead of the holistic one. As in the case of the general impression scoring method, the holistic rubric assessment is not undertaken on the basis of each dimension, but only one score is given regarding the overall performance. Therefore,

268

it can be said that the research findings of the studies carried out by Alharby (2006) and Wiseman (2008, 2012) parallel the findings of the present study. However, it would be more accurate to define this parallelism as a partial similarity instead of a complete correspondence. Although holistic rubric scoring and scoring via the general impression share a commonality—in that there is only one score given through the consideration of overall performance—other qualities of these two methods differentiate them from one another. Concerning the holistic rubric, performance is not divided into sub-components; rather, different levels are defined for general performance. The rater takes the related definitions as a base while deciding upon the level to which the performance/product of the student corresponds. On the other hand, there no such criteria can be found in general impression scoring, only the interval in which the scoring will is to be undertaken is certain. The rater makes the assessment according to the performance definitions made by themselves within the prescribed interval. It is obvious that, even though both holistic rubric and scoring by means of general impression have different qualities, neither can give feedback as detailed and rich as that given using the analytic rubric.

When the results reported for the peer rater facet at the end of the many-facet Rasch analysis are examined, a statistically significant difference between the raters in terms of their severity in the analytic rubric scoring was found. On the other hand, it was observed that the peer raters gave similar scores in the general impression scoring. Essentially many studies in the literature (Chi, 2001; Knoch, 2009; Ghalib & Al-Hattami, 2015; Jönsson & Balan, 2018) suggest that the analytic rubric is expected to create a common frame of reference among the raters, thereby improves rater reliability. Consequently, it is surprising that, in the current study no statistically significant difference was found among raters regarding general impression scoring, whereas a statistically significant difference was found among raters regarding analytic rubric. Despite this, other studies in the literature that have reached findings that parallel those of the current study. For example, Çetin and Kelecioğlu (2004) carried out a study investigating the relationships between scores via scoring key and general impression in essay type exams. This study found that inter-rater reliability of the general impression scoring was higher compared with scoring key. The research findings of the study carried out by Ounis (2017) also parallel those of the current study. Ounis (2017) compared analytic rubric and holistic rubric scoring in order to designate which one was better when assessing speaking skills, and concluded that agreement between raters was higher when they used the holistic rubric. However, it should be noted that, at this point, holistic rubric scoring and scoring via general impression are different scoring methods and so the conformity between of Ounis's (2017) research findings and the findings of the current study do not extend beyond an indirect similarity.

While statistically significant differences were found between the peer raters in the analytic rubric scoring, no such a difference was determined when scoring is undertaken via general impression. This can be explained as follows: Only one score is given for each student in scoring via general impression. On the other hand, when analytic rubric is used, the 13 dimensions are scored separately, meaning that each student receives 13 times as many scores as part of the analytic rubric as compared with general impression scoring. It is possible for a difference to exist between raters for each scoring method. When the analytic rubric is used, the increase in the number of scores might also have increased the possibility of observing differences among raters. Consequently, it may be more appropriate to interpret the results in regard to peer raters in consideration of the aforementioned possibility.

In this study it was concluded that a positive high correlation exists between ability estimations calculated according to scores given by the peers, and by the instructor for both via analytic rubric and general impression scoring methods. According to this, the scores given by the peers conform with those scores given by the instructor regarding both methods. This result parallels the research findings of other studies in the literature (Alzaid, 2017; Napoles, 2008; Şahin, 2008). Falchikov and Goldfinch (2000) carried out a meta-analysis using 48 quantitative peer assessment studies comparing peer and teacher marks; they found that the mean correlation between the marks given by peers and the instructor was .69 across all the studies within the meta-analysis.

Another finding of the study is that ability estimations, corresponding to scoring via peer raters' analytic rubric and general impression scores, held a positive and highly strong relation. This finding revealed that in the both scoring methods, students are ranked in a similar way in terms of their ability levels. In the light of this finding, it can be said no significant difference in ranking would be seen between the use of analytic rubric scoring and general impression scoring for those assessments aiming to rank the examinees in terms of their ability. This research finding corroborates that "*there is a high correlation between ability estimations calculated according to different scoring methods*" in studies carried out with teachers/instructors/expert raters, and that focusing on comparing analytical and holistic rubrics (Anita, 2011; Chi, 2001; Ghalib & Al-Hattami, 2015; Hunter, Jones, Radhawa, 1996; Yune, Lee, Im, Kam, & Baek, 2018) is also valid for the scoring via analytic rubric and general impression in peer assessment.

Although it is not directly related to the problems sought in the research, another finding reached in the scope of the analyses conducted within this study is that there is no halo effect in the analytic rubric scoring undertaken by peer raters. Halo effect manifests itself as the inability to distinguish among the dimensions of the analytical rubric. The fact that reliability coefficient and separation index regarding the dimension facet are high—and that the chi-square value was found to be significant when scoring using the analytic rubric—shows that students' performances in different dimensions of the rubric can be scored independently one another. This also gives a clue that halo effect is not involved in scoring. Although halo effect is the most common rater error in scoring when using the analytic rubric (Myford & Wolfe, 2004), this error did not show up among scoring by peer raters; this is an important point in alleviating the concerns about the accurate use of analytic rubric in peer assessment.

Based on the results of the research, it can be said that both analytical rubric and general impression scoring can be used in peer assessment. However, the results of the research may have been influenced by the fact that the scorings were undertaken according to the analytical rubric firstly and subsequently the general impression. More clearly; despite the preventions taken, the scoring done via analytical rubric may have influenced the scoring based on the general impression. Therefore, in future research on the subject, it can be examined whether changing the ranking of scoring methods will make a difference in obtained results. In fact, a research design can be created in which half of the peer raters first score based on the general impression and the other half score first via the analytical rubric.

REFERENCES

Alharby, E. R. (2006). *A comparison between two scoring methods, holistic vs. analytic using two measurement models, generalizability theory and the many facet Rasch measurement within the context of performance assessment.* (Unpublished doctoral dissertation). Pennsylvania State University, Pennsylvania.

Alzaid, J. M. (2017). The effect of peer assessment on the evaluation process of students. *International Education Studies*, *10*(6), 159–173. DOI: 10.5539/ies.v10n6p159

Amo, E., & Jareno, F. (2011). Self, peer and teacher assessment as active learning methods. *Research Journal of International Studies*, *18*, 41–47.

Anita, H. (2011). The connection between analytic and holistic approaches to scoring in the writing component of the PROFEX EMP exam. *Acta Medica Marisiensis*, *57*(3), 206–208.

Ashraf, H., & Mahdinezhad, M. (2015). The role of peer-assessment versus self-assessment in promoting autonomy in language use: A case of EFL learners. *Iranian Journal of Language Testing, 5*(2), 110–120.

Berry, R. (2008). *Assessment for learning.* Hong Kong: Hong Kong University Press.

Bostock, S. (2000). Student peer assessment. Retrieved from https://www.reading.ac.uk/web/files/engageinassessment/student_peer_assessment_-_stephen_bostock.pdf

Boud, D. (2013). Introduction: making the move to peer learning. Boud, D., Cohen, R., & Sampson, J. (Eds.), in *Peer learning in higher education learning from & with each other* (pp. 1–18). New York: Routledge.

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many–facet Rasch model. *Journal of Applied Measurement*, *2*(4), 379–388.

Çetin, B., & Kelecioğlu, H. (2004). The relation between scores predicted from structured features of essay and scores based on scoring key and overall impression in essay type examinations. *Hacettepe University Journal of Education*, *26*, 19–26.

Falchikov, N. (2001). *Learning together: Peer tutoring in higher education*. London: Routledge Falmer.

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. DOI: 10.3102/00346543070003287

Frankland, S. (2007). Peer assessments among students in a problem based learning format. In S. Frankland (Eds.), *Enhancing teaching and learning through assessment: Developing an appropriate model* (pp. 144–155). Dordrecht: Springer.

Ghalib, T. K., & Al-Hattami, A. A. (2015). Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching*, *8*(7), 225–236. DOI: 10.5539/elt.v8n7p225

Gravells, A. (2014). *The award in education and training*. London: Learning Matters.

Gronlund, N. E. (1998). *Assessment of student achievement*. Boston: Allyn & Bacon.

Güler, N., İlhan, M., Güneyli, A., & Demir, S. (2017). An evaluation of the psychometric properties of three different forms of Daly and Miller's writing apprehension test through Rasch analysis. *Educational Sciences: Theory & Practice*, *17*(3), 721–744. DOI: 10.12738/estp.2017.3.0051

Hall, E. K., & Salmon, S. J. (2003). Chocolate chip cookies and rubrics helping students understand rubrics in inclusive settings. *Teaching Exceptional Children*, *35*(4), 8–11. DOI: 10.1177/004005990303500401

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications, Inc.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self– and peer–assessment: The students' views. *Higher Education Research & Development*, *20*(1), 53–70. DOI: 10.1080/07294360123776

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1979). *Applied statistics for the behavioral sciences*. Chicago: Rand McNally College.

Hunter, D. M., Jones, R. M., Radhawa, B. S. (1996). The use of holistic versus analytic scoring for large–scale assessment of writing. *The Canadian Journal of Program Evaluation*, *11*(2), 61–85.

Jönsson, A., & Balan, A. (2018). Analytic or holistic: A study of agreement between different grading models. *Practical Assessment, Research & Evaluation*, *23*(12), 1–11.

Kan, A. (2007). An alternative method in the new educational program from the point of performance-based assessment: Rubric scoring scales. *Educational Sciences: Theory & Practice*, *7*(1), 144–152.

Karaca, E. (2009). An evaluation of teacher trainees' opinions of the peer assessment in terms of some variables. *World Applied Sciences Journal, 6*(1), 123–128.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(20), 275–304. DOI: 10.1177/0265532208101008

Lee, M., Peterson, J. J., & Dixon, A. (2010). Rasch calibration of physical activity self-efficacy and social support scale for persons with intellectual disabilities. *Research in Developmental Disabilities*, *31*(4), 903−913. DOI: 10.1016/j.ridd.2010.02.010

Lester, F. K., Lambdin, D. V., & Preston, R. V. (1997). A new vision of the nature and purposes of the assessment in the mathematics classroom. Phye, G. D. (Eds.) in *Handbook of Classroom Assessment: Learning, Adjustment, and Achievement* (pp. 287−320). San Diego, CA, US: Academic Press.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85−106.

Linacre, J. M. (2018). A user's guide to FACETS Rasch-model computer programs. Retrieved from https://www.winsteps.com/a/Facets-Manual.pdf

Liu, N. F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 279−290. DOI: 10.19080/13562510600680582

Macpherson, K. (1999). The development of critical thinking skills in undergraduate supervisory management units: Efficacy of student peer assessment. *Assessment & Evaluation in Higher Education*, *24*(3), 273−284. DOI: 10.1080/0260293990240302

Mann, B. L. (2006). Testing the validity of the post and vote model of web-based peer assessment. In D. D. Williams, S. L. Howell, & M. Hricko (Eds.), *Online assessment, measurement, and evaluation: Emerging practices* (pp. 131−152). Hershey, PA: Information Science.

Matsuno, S. (2009). Self-, peer–, and teacher–assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75−100. DOI: 10.1177/0265532208097337

Mehrdad, N., Bigdelib, S., & Ebrahimia, H. (2012). A comparative study on self, peer and teacher evaluation to evaluate clinical skills of nursing students. *Procedia - Social and Behavioral Sciences*, *47*, 1847−1852.

McDonald, B. (2016). *Peer assessment that works: A guide for teachers*. London: Rowman & Littlefield.

271

McLeod, S. G., Brown, G. C., McDaniels, W., & Sledge, L. (2009). Improving writing with a PAL: harnessing the power of peer assisted learning with the reader's assessment rubrics. *International Journal of Teaching and Learning in Higher Education*, *20*(3), 488−502.

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

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation, 7*(3), 1-5. Retrieved from: https://pareonline.net/getvn.asp?v=7&n=3

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10), 1-6. Retrieved from http://pareonline.net/getvn.asp?v=7&n=10

Mutwarasibo, F. (2016). University students' attitudes towards peer assessment and reactions to peer feedback on group writing. *Rwanda Journal, Series A: Arts and Humanities*, *1*(1), 32−48. DOI: 10.4314/rj.v1i1.4A

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.

Napoles, J. (2008). Relationships among instructor, peer, and self-evaluations of undergraduate music education majors' micro-teaching experiences. *Journal of Research in Music Education*, *56*(1), 82−91. DOI: 10.1177/0022429408323071

Nitko, A.J. (2004). *Educational assessment of students*. Upper Saddle River, NJ: Pearson.

Ounis, M. (2017). A comparison between holistic and analytic assessment of speaking. *Journal of Language Teaching and Research*, *8*(4), 679−690. DOI: 10.17507/jltr.0804.06

Petkov, D., & Petkova, O. (2006). Development of scoring rubrics for IS projects as an assessment tool. *Issues in Informing Science and Information Technology*, *3*, 499−510. DOI: 10.28945/910

Prins, F. J., Sluijsmans, D. M. A., Kirschner, P. A., & Strijbos, J. W. (2005). Formative peer assessment in CSCL environment: a case study. Assessment & Evaluation in Higher Education, *30*(4), 417−444. DOI: 10.1080/02602930500099219

Reddy, M. Y. (2010). Design and development of rubrics to improve assessment outcomes. A pilot study in a master's level business program in India. *Quality Assurance in Education*, *19*(1), 84−104. DOI: 10.1108/09684881111107771

Sluijmans, D. M. A., Brand-Gruwel, S., van Merrienboer, J. J. G., & Bastiaens, T. J. (2003). The training of peer assessment skills to promote the development of reflections skills in teacher education. *Studies in Educational Evaluation*, *29*(1), 23−42. DOI: 10.1016/S0191-491X(03)90003-4

Smith, H., Cooper, A., & Lancaster, L. (2002) Improving the quality of undergraduate peer assessment: A case for student and staff development. *Innovations in Education and Teaching International*, *39*(1), 71−81. DOI: 10.1080/13558000110102904

Şahin, M. G., Taşdelen Teker, G. & Güler, N. (2016). An analysis of peer assessment through many facet Rasch model. *Journal of Education and Practice*, *7*(32), 172−181.

Şahin, S. (2008). An application of peer assessment in higher education. *The Turkish Online Journal of Educational Technology*, *7*(2), 5−10.

Şaşmaz Ören, F. (2018). Self, peer and teacher assessments: What is the level of relationship between them? *European Journal of Education Studies, 4*(7), 1−19. DOI: 10.5281/zenodo.1249959

Tan, Ş. (2015). *Eğitimde ölçme ve değerlendirme KPSS el kitabı [Measurement and evaluation in education: KPSS handbook]*. Ankara: Pegem.

Taşdelen Teker, G., Şahin, G., & Baytemir, K. (2016). Using generalizability theory to investigate the reliability of peer assessment. *Journal of Human Sciences*, *13*(3), 5574−5586. DOI: 10.14687/jhs.v13i3.4155

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

Topping, K. J. (2009) Peer assessment. *Theory into Practice*, *48*(1), 20−27. DOI: 10.1080/00405840802577569

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. DOI: 10.1080/713611428

van den Berg, I., Admiraal, W., & Pilot, A. (2006). Designing student peer assessment in higher education: analysis of written and oral peer feedback. *Teaching in Higher Education*, *11*(2), 135−147. DOI: 10.1080/13562510500527685

Wen, M. L., & Tsai, C. C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education*, *51*(1), 27−44. DOI: 10.1007/s10734-004-6375-8

Wiseman, C. S. (2008). *Investigating selected facets in measuring second language writing ability using holistic and analytic scoring methods* (Unpublished doctoral dissertation). Columbia University. New York.

Wiseman, C. S. (2012). Comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *Iranian Journal of Language Testing*, *2*(1), 59−92.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370–371.

272

Yakar, L. (2019). Tamamlayıcı ölçme ve değerlendirme teknikleri III [Complementary measurement and evaluation techniques]. N. Doğan (Eds.), in *Eğitimde Ölçme ve Değerlendirme [Measurement and Evaluation in Education]* (pp. 245−270). Ankara: Pegem.

Yune, S. J., Lee, S. Y., Im, S. J., Kam, B. S., & Baek, S. Y. (2018). Holistic rubric vs. analytic rubric for measuring clinical performance levels in medical students. *BMC Medical Education*, *18*(1), 1−6. DOI: 10.1186/s12909-018-1228-9

273

TÜRKÇE GENİŞLETİLMİŞ ÖZET

Günümüzde yaşam boyu öğrenme becerisine sahip, bilgi ve teknoloji çağının öngördüğü beceriler ile donanmış bireylerin yetişmesine katkı sağlayacak değerlendirme sistemlerinin diğer eğitim kademlerinde olduğu gibi yükseköğretimde de kullanılması ve bilhassa geleceğin öğretmenlerini yetiştiren kurumlar olması sebebiyle eğitim fakültelerinde kendi yerini bulması gerekmektedir. Bu bağlamda; genelde yükseköğretimde özelde ise eğitim fakültelerinde, 21. yüzyıl becerilerinin bireylere kazandırılmasına yardımcı olacak değerlendirme yaklaşımlarına başvurulmasının bir ihtiyaç olduğu söylenebilir. Bahsedilen ihtiyaca cevap olabilecek yaklaşımların başında, öğrenci değerlendirmelerini ölçme-değerlendirme sürecine katan yöntemler gelir ki; bunlardan biri akran değerlendirmedir. Dolayısıyla yükseköğretim düzeyinde akran değerlendirmeyi farklı yönleriyle ele alan, hangi şekilde kullanıldığında daha işlevsel olacağını ve daha doğru sonuçlar vereceğini ortaya koyan çalışmaların yapılması oldukça önemlidir. Bu kapsamda çalışmada, yükseköğretim öğrencilerinden oluşan bir örneklem üzerinde; akran değerlendirmede analitik rubrikle ve genel izlenimle yapılan puanlamaların karşılaştırılması amaçlanmaktadır. Bu doğrultuda, *i)* akranların analitik rubrikle ve genel izlenimle yaptıkları puanlamalara ilişkin güvenirlik değerlerinin belirlenmesi, *ii)* iki puanlama yöntemi için ayrı ayrı olmak üzere akranların yaptığı puanlamalara göre hesaplanan yetenek kestirimleri ile dersi yürüten öğretim elemanın yaptığı puanlamalardan elde edilen yetenek ölçüleri arasındaki korelasyon katsayılarının incelenmesi ve *iii)* akranların analitik rubrikle ve genel izlenimle yaptıkları puanlamalara karşılık gelen yetenek kestirimleri arasındaki tutarlılığın test edilmesi hedeflenmektedir.

Araştırmanın çalışma grubunu, 2018-2019 eğitim öğretim yılında Türkiye'de bir devlet üniversitesinin Okul Öncesi Öğretmenliği Anabilim dalı üçüncü sınıfında öğrenim gören 66 öğrenci oluşturmaktadır. Bu öğrencilerden altısı gönüllülük esasına dayalı olarak akran değerlendirmeyi gerçekleştirmek üzere seçilmiştir. Çalışma, bilimsel araştırma yöntemleri dersinde yürütülmüştür. Öğrencilerden ilgili ders kapsamında örnek bir çalışma hazırlamaları ve hazırladıkları çalışmayı sınıf ortamında sunmaları istenmiştir. Öğrencilerin sunum yaptıkları esnada, akranlar ve dersin öğretim elemanı her bir öğrenci için önce analitik rubrik kullanarak ve sonrasında genel izlenimle puanlama yapmıştır. Puanlamadan elde edilen veriler Rasch modeline göre analiz edilmiştir. Araştırmanın birinci alt problemi doğrultusunda; analitik rubrikle ve genel izlenimle yapılan puanlamalarda puanlayıcı ve birey yüzeylerine ait güvenirlik, ayırma indeksi ve Ki Kare değerleri karşılaştırmalı olarak incelenmiştir. İkinci alt problem kapsamında; iki puanlama yöntemi için ayrı ayrı olmak üzere akranların yaptığı puanlamalara göre hesaplanan yetenek kestirimleri ile dersi yürüten öğretim elemanın yaptığı puanlamalardan elde edilen yetenek ölçüleri arasındaki korelasyon katsayısı (Pearson momentler çarpımı korelasyonu) hesaplanmıştır. Benzer şekilde; araştırmanın üçüncü alt problemi için korelasyon analizi uygulanarak akranların analitik rubrikle ve genel izlenimle yaptıkları puanlamalara karşılık gelen yetenek kestirimleri arasındaki tutarlılık test edilmiştir. Çalışmada, Rasch analizleri için FACETS 3.70.1 paket programı kullanılırken; korelasyon analizleri SPSS 21.0 paket programında gerçekleştirilmiştir.

Araştırmada, her iki puanlama yönteminde de bireylerin yüksek güvenirlikte birbirinden ayırt edildiği belirlenmiştir. Bununla birlikte, analitik rubrik kullanıldığında bireylerin yetenek düzeyleri arasındaki farklılıkların daha hassas bir biçimde ortaya konulduğu saptanmıştır. Buna göre, öğrencilerin yetenek düzeyleri arasındaki küçük farklılıkların alınacak kararları değiştirilebildiği bir değerlendirme yapılıyorsa puanlamanın genel izlenimle değil de analitik rubrik kullanılarak gerçekleştirilmesi daha doğru bir tercih olacaktır. Küçük farklılıkların bireylerin yetenek düzeylerine ilişkin sıralamaları etkilemeyeceği bir değerlendirme yapılması durumunda ise daha ekonomik olması bakımından genel izlenimle puanlama tercih edilebilir. Çalışmada hem analitik rubrikle hem de genel izlenimle yapılan değerlendirmede; akranlar ile öğretim elemanının verdiği puanlar üzerinden hesaplanan yetenek

274

kestirimleri arasında, pozitif yönlü yüksek korelasyonlar bulunmuştur. Buna göre; her iki puanlama yöntemi için de akranların yaptığı puanlamalar, öğretim elemanın verdiği notlar ile tutarlıdır. Bu sonuç alanyazındaki araştırmalar ile örtüşmektedir.

Akranların analitik rubrikle ve genel izlenimle yaptıkları puanlamalara karşılık gelen yetenek kestirimlerinin pozitif yönlü güçlü bir ilişki içerisinde olması, çalışmada ulaşılan bir diğer sonuçtur. Bu sonuç; her iki yönteme göre yapılan puanlamalarda, bireylerin yetenek düzeyleri açısından büyük ölçüde benzer şekilde sıralandığını göstermektedir. Bu bulguya dayanarak bireyleri yetenek düzeyleri açısından sıralamak amacıyla yapılan bir değerlendirmede, analitik rubriğe ya da genel izlenime göre puanlama yapılmasının sıralamalar arasında ciddi bir fark yaratmayacağı ifade edilebilir. Araştırmaya ilişkin bu sonuç, öğretmenler/öğretim elemanı/uzman puanlayıcılar üzerinde yürütülen ve analitik ile holistik rubriğin karşılaştırılmasına odaklanan çalışmalardan elde edilen "*farklı puanlama yöntemlerine göre hesaplanan yetenek ölçüleri arasında yüksek bir korelasyon olduğu*" bulgusunun akran değerlendirme ve genel izlenimle puanlama için de geçerli olduğunu ortaya koymaktadır.

Doğrudan araştırmada yanıt aranan problemler ile ilgili olmasa da çalışma kapsamında yapılan analizler doğrultusunda varılabilecek bir diğer sonuç, akranlar tarafından yapılan puanlamalarda analitik rubrikteki boyutların birbirinden ayırt edilememesi şeklinde kendisi gösteren halo etkisine rastlanmadığı şeklindedir. Analitik rubrik ile yapılan puanlamalarda madde yüzeyine ilişkin güvenirlik katsayısı ile ayırma indeksinin yüksek ve Ki Kare değerinin anlamlı çıkması, öğrencilerin rubriğin farklı boyutlarındaki performanslarının birbirinden bağımsız olarak puanlanabildiği göstermekte ve yapılan puanlamalara halo etkisinin karışmadığına dair ipucu vermektedir. Halo etkisi, analitik rubrik ile yapılan puanlamalarda en sık karşılaşılan puanlayıcı hatası olmasına karşın, akranlarca yapılan puanlamalarda bu hatanın ortaya çıkmaması akran değerlendirmede analitik rubriğin ne kadar doğru kullanılabileceği yönündeki endişeleri azaltması bakımından önemlidir.

Araştırmada ulaşılan sonuçlardan hareketle, akran değerlendirmede hem analitik rubriğe hem de genel izlenime dayalı puanlamanın yapılabileceği söylenebilir. Bununla birlikte; çalışmadan elde edilen sonuçlar, puanlamaların önce analitik rubriğe ve ardından genel izlenime göre yapılmasından etkilenmiş olabilir. Dolayısıyla konu hakkında yapılacak ileri araştırmalarda puanlama yöntemlerinin sıralamasının değiştirilmesinin ulaşılan sonuçlarda fark yaratıp yaratmayacağı incelenebilir.

275