ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 62-75

Examining The Rater Drift in The Assessment of Presentation Skills in Secondary School Context

Aslıhan ERMAN ASLANOĞLU* Mehmet ŞATA**

Abstract

The alternative assessment, including peer assessment, helps students develop metacognition among the sub-categories of assessment types. Despite the advantage of alternative assessment, reliability and validity issues are the most significant problems in alternative assessment. This study investigated the rater drift, one of the rater effects, in peer assessment. The performance of 8 oral presentations based on group work in the Science and Technology course was scored by 7th-grade students (N=28) using the rubric researchers developed. The presentations lasted for four days, with two presentations each day. While examining the time-dependent drift in rater severity in peer assessment, the many-Facet Rasch Measurement model was used. Two indexes (interaction term and standardized differences) were calculated with many-facet Rasch measurement to determine the raters who made rater drift either individually or as a group. The analysis examined the variance of scores in the following days compared to the first day's scores. Accordingly, the two methods used to determine rater drift gave similar results, and some raters at the individual level tended to be more severe or lenient over time. However, no significant rater drift at the group level showed that drifts had no specific models.

Keywords: Alternative assessment, many-facet Rasch measurement, peer assessment, validity, rater drift

Introduction

There has been an increase in the demand for qualified labor over the past century. Occupational groups desire individuals who can solve problems, think critically, analyze and present data effectively, have effective verbal and written communication skills, and evaluate themselves and their peers (Dochy, 2001). Education plays a crucial role in raising individuals; therefore, raising individuals with such characteristics can be accomplished through education systems oriented in this direction (Batmaz et al., 2022; Kaya et al., 2023). However, traditional assessment approaches applied in the learning environment are insufficient in measuring the mentioned characteristics. This new understanding necessitates establishing a connection between learning and assessment processes, which heightens the use of alternative assessment in education (Oosterhof, 2003).

Unlike traditional approaches, students are not just passive recipients of information in alternative assessment approaches. The most significant characteristic of these approaches is to make individuals develop higher-order thinking skills, such as critical and creative thinking and problem-solving skills, by actively participating in the process (Kutlu et al., 2014). Having gained importance with new approaches, performance evaluation measures how well the student uses the basic knowledge they have gained while performing complex tasks in real life. In this respect, performance assessment is unlike classical tests (multiple-choice, short answer, matching, etc.) which are concerned with the student's ability to retrieve information from memory as it is. It is based on the process of actively constructing knowledge and using it in real life. (Moore, 2009). Students should be allowed to interact with their peers and teachers during this process. Thus, it becomes possible for students to construct knowledge and share structured knowledge. Assessment and evaluation are indispensable components of learning.

* Asst. Prof. Dr., Ufuk Universty, Faculty of Education, Ankara-Türkiye, aslihanerman@yahoo.com, ORCID ID: 0000-0002-1364-7386
** Asst. Prof. Dr., Ağrı İbrahim Çeçen University, Faculty of Education, Ağrı-Türkiye, mehmetwsata@gmail.com, ORCID ID: 0000-0003-2683-4997

To cite this article:
Erman-Aslanoğlu, A., & Şata M. (2023). Examining the rater drift in the assessment of presentation skills in secondary school context. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 62-75. https://doi.org/10.21031/epod.1213969

Received: 3.12.2022
Accepted: 15.03.2023

They positively affect the learning process by improving the quality of learning and improving the learners' sense of thinking and autonomy (Orsmond et al., 2000). Alternative assessment, which includes peer assessment, has been accepted as the one that allows students to develop metacognitive skills among the subcategories of assessment types (Liu & Brantmeier, 2019).

The peer assessment emphasizes formative purposbushees to show students' performance rather than the summative purposes of assessment by enabling students to take responsibility for their learning (Azarnoosh, 2013). Studies assert that students learn better when they can benefit from the opinions of their peers (Black et al., 2003; Topping, 2017). Peer assessment can also be one of the guiding elements of group work appropriately conducted, which is essential in today's business life. In this respect, peer assessment practice carried out in group work can contribute to the success of individuals as it can increase their responsibility of individuals (Falchikov & Goldfinch, 2000; Yurdabakan & Cihanoğlu, 2009). In addition, students have versatile feedback on the quality of their work to the extent of a single instructor evaluation which is much more classical than peer assessment (Topping, 2009).

When peer assessment is used in the teaching process, the most important problem is the reliability and validity of the scores obtained from these sources (Donnon et al., 2013; Hafner & Hafner, 2003; Topping, 2003). A limited number of studies on the reliability and validity of peer assessment emphasize the importance of peer assessment in the teaching process and state that its validity and reliability are sufficient when appropriately done (Patri, 2002). In order to increase the reliability of the results obtained from peer reviews, it is of great importance to try to increase rater reliability. Validity of the results obtained from the performance measurement is required to ensure the scores' reliability (Hafner & Hafner, 2003; Jonsson & Svingby, 2007).

However, while evaluating students' performance, different factors arising from the raters can interfere with the measurement results. Rater-based factors affecting student performance are called rater effects (Farrokhi et al., 2011). Errors in rater decisions (i.e., rater effects) can influence the accuracy of assigned ratings. Although there are various errors originating from the raters in the performance assessment process, the most common rater-effective errors in the literature are rater severity and leniency and are discussed as halo effect, central tendency, range restriction, and differential rater functioning (DRF) (Myford & Wolfe, 2003). Another rater effect type that has recently attracted researchers' attention is the differential rater functioning over time (drift). Accordingly, the issue of whether the rater effect changes over time becomes significant with the increase in the use of large-scale tests, especially the implementation and scoring of which spread over time, and it has been tried to determine whether rater behaviors are affected by this situation and cause measurement errors (Congdon & McQueen, 2000; Harik et al., 2009; Myford & Wolfe, 2009).

Drift is the changes that occur in the scoring of students' performance at different times and in the rater behaviors depending on time (Wolfe et al., 2007). Various types of drift have been defined concerning the direction in which drift manifests itself. Recency drift and primacy drift are the most common (Myford & Wolfe, 2003). In Primacy drift, raters give high scores when scoring and tend to give lower scores as the rating progresses, yet Recency drift refers to the opposite. In summary, in this common type of drift, raters may display more severe or lenient behavior over time (differential severity). However, the item facet -calibration should be performed - must remain constant so that the measurement results can be compared in scoring at different times. (Leckie & Baird, 2011). Especially in cases where more than one rater implements the scoring process (i.e., peer assessment), rater calibration is not to change from person to person and from time to time (Congdon & McQueen, 2000). Such situations threaten the validity of scores by causing irrelevant variance related to students' performance (Messick, 1994).

Previous studies made several suggestions in order to reduce rater-related errors, including training raters (Hauenstein & McCusker, 2017; Knoch et al., 2007), involving more than one rater in the process (Kubiszyn & Borich, 2013), using rubrics (Andrade, 2005; Oosterhof, 2003), and adding such methods in the classroom more often (Bushell, 2006; Topping, 2003). As such, there can be less concern about the reliability of the scores. Researchers also recommended the Multi Facet Rasch Model (MFRM) to determine the reliability and validity of peer review scores (Farrokhi et al., 2011; Myford & Wolfe, 2009). MFRM removes the limitations of Classical Test Theory approaches. In evaluating students' performance in MFRM, the factors that may affect the scores of the students are not limited to the ability levels of the individuals or the difficulty levels of the items used in the measurement process. Factors related to raters can also cause variability in student performance scores (Baird et al., 2013), which makes MFRM a suitable option for performance evaluations affected by rater behavior (Mulqueen et al., 2000).

It has been observed that the limited number of studies on drift in the literature are generally related to higher education level and second language English teaching proficiency exams. The findings of these studies, which tried to determine whether the scores given by the raters changed over time, showed different results. Some studies found rater drifts over time (Börkan, 2017; Congdon & McQueen, 2000; McLaughlin et al., 2009; Myford, 1991; Myford & Wolfe, 2009; Pinot de Moira et al., 2002). For example, Myford (1991) found rater severity due to evaluating students' drama performance for a month by referees with different experiences. Another study using MFRM determined the severity estimates of 10 raters who scored elementary school students' English writing tasks using rubrics for seven days (Condon & Mcquenn, 2000). According to the research results, it has been observed that while 9 of the raters gave more severe scores over time, one rater started to give more tolerant scores. As a result, these researchers have found rater severity for raters daily, but there was no general pattern of change. On the other hand, some researchers have not encountered scoring severity as a result of their studies (Humphris & Kaney, 2001; Leckie & Baird, 2011). For instance, Humphris and Kaney (2001) have concluded that in a 4-station exam (all stations took 5 minutes with a simulated patient) to measure the communication skills of first-year undergraduate medical students with the patient, the raters did not make rater drifts during the scoring made at different times. Similarly, in their study, Leckie and Baird (2011) concluded that the raters evaluating the scores of 14-year-old students from a large-scale English writing skill test did not make rater drifts. Some researchers, on the other hand, have obtained different results according to the structure of the exams related to rater severity. For instance, Lunz and Stahl (1990), in their study using MFRM, which took four days to score, investigated whether the raters in the oral exam, English composition exam, and clinical exams made rater drift in the scoring that lasted for four days. Research findings have shown that while rater drifts were observed in clinical and English composition exams, raters did not make any time-dependent rater drifts in the oral exam.

Purpose and Significance of the Research

Most peer assessment studies have been carried out at the higher education level, especially in foreign language teaching. The compatibility of peer scores with teacher scores is the basis for determining validity and reliability. These studies are based on the assumption that teacher scores are valid and reliable. However, teacher scores may not always be reliable and may be affected by various errors. To increase the reliability of the results obtained from peer reviews, it is of great importance to try to increase rater reliability because the validity of the results obtained from the performance measurement is possible with the reliability of the scoring (Jonsson & Svingby, 2007). Prejudices in assessment results arising from raters for different reasons threaten validity as they are sources of variance unrelated to the construct (Messick, 1994). Therefore, it is necessary to provide evidence of the validity of peer review scores. However, the evidence for the validity of peer assessment in the literature is limited (Börkan, 2017). In addition, using MFRM on rater effects provides statistical approaches to identify and correct some of these rater biases in the studies. While these approaches do not guarantee that all rater effects will be identified and deleted from test scores, they provide important information about identifying and taking action on a significant portion of rater effects.

In the present study, not only the rubrics but also the multiple raters were used in the process of measuring students' performance in peer assessment in order to satisfy the reliability of the measurement, and the MFRM approach was used.

Considering all of these, the current study attempts to explore the status of the rater effect in performance scoring when peer assessment was extended to a process. In this regard, the study, implemented with secondary school students, aims to reveal the students' behaviors in the peer-assessment process spreading over four separate days with the help of the rubrics. In particular, the study seeks to answer the following questions:

1. Do the raters as a group demonstrate differential severity/leniency behavior throughout the scoring period?

2. Does any individual rater demonstrate a severity/leniency interaction throughout the scoring period?

Methods

Research Model

This study aims to examine the changes in the ratings of peer raters over time in evaluating the presentation skills of 7th-grade students. For this purpose, as one of the quantitative research approaches, the descriptive research model was used (Şata, 2020).

The Study Group

The present study was conducted in the first semester of the 2021-2022 education year in Turkiye. The study group of the research consisted of 7th-grade students (N=28) studying in a private school in Ankara, Çankaya district. The presentations of the science and technology course prepared by the 7th-grade students in groups of three (i.e., eight groups) were scored by peer raters using rubrics. Since each presentation was made to a group of three people, the total number of peer raters was 25, and 8 groups carried out the scoring.

Instruments

In peer assessment of presentations, students use an analytical rubric developed by the researchers (see Appendix 1). While developing the rubric, researchers determined the criteria for assessing students' presentation skills by reviewing the relevant literature. The scale was developed as an analytical rubric, preferable to the holistic one since they provide more detailed feedback on student performance. One of the most important advantages of analytical rubrics is that they are better than holistic rubrics in providing both intra-rater and inter-rater reliability in the assessment process (Andrade, 2005), which is the main reason that led us to adopt them in the present study. The response categories were initially developed as five but were subsequently reduced to four. Respectively, content, coherence, use of material, communication, and use of time. To provide evidence for the reliability and validity of the measurements obtained from the analytical rubric, content validity rates and Kendall Tau coefficient were calculated through expert opinions. In line with the opinions of field experts (6 people) and assessment and evaluation experts (3 people), the rating of the rubric was also reduced from five to four in the final version. To measure the reliability, Kendall tau was calculated to measure the relation between the scores of the two randomly selected groups. It was determined that the rubric had an acceptable reliability score (r = .652; p < .001).

Data Collection

Since the present research attempts to examine rater change over time, eight groups made presentations, two groups each day, in four days. In the evaluation process of the presentations, the analytical rubric was used, and the criteria of the measurement tool were introduced to the students one by one before they made their presentations to ensure the consistency of the measurements. The scores of the peer raters at the end of each presentation were collected. The next group made their presentation after having a short break. Students followed the same procedure on each of the four days.

Data Analysis

Statistical analysis of the research was carried out using the multi-facet Rasch model, which is a member of the Rasch model family. Many-Facets Rasch Analysis (Many-Facets Rasch Model) is a useful measurement model since it considers all sources of variability that affect individuals' performance or skill levels (Baird et al., 2013; Kim et al., 2012; Linacre, 2017). In addition, it provides the opportunity to examine the interaction among the variability sources (Kassim & Noor, 2007). The simultaneous analysis of both two- and multi-category measurements increases the applicability of the multi-facet Rasch model. In the current study, the Many-Facets Rasch model was used since it aimed to examine both the main effects and the common interactions among sources of variability in the peer-assessment process. In this study, standardized differences (Signed Area Index, SAI) obtained from multi-facet Rasch analysis and interaction terms were used to examine the change of peer raters over time.

Measurements at different times were estimated as separate models to explore standardized differences. Estimated logit values for each model were divided by standard errors, and standardized values were obtained. In this study, since student presentations were different in the measurements made at different times, the mean score of the estimations of the raters was modeled to be zero to eliminate the influence of this difference. As such, the relative change in the strictness or leniency behaviors of the raters could be examined in the scoring that takes place at different times. For this reason, student presentations (groups) were handled as non-center in the research. For the standardized differences, the first measurement baseline time was taken, the measurements at other times were compared with the baseline time, and score deviations (SAI) were calculated. The SAI value has been standardized by the formula given below.

$$Z_{SAI_{Difference}} = \frac{M_c - M_b}{\sqrt{SE^2_{M_c} + SE^2_{M_b}}} \qquad (1)$$

Here, Mc corresponds to rater strictness or leniency compared to baseline time, while Mb corresponds to rater strictness and leniency at baseline time. The two values in the denominator represent the squares of the standard errors of the rater severity or leniency values at two different times. If the $Z_{SAI_{Difference}}$ value is calculated from two different times out of the range of ±1.96 values, it indicates a statistically significant difference (Raju, 1990). When the direction of the facets is positive, positive $Z_{SAI_{Difference}}$ values are interpreted as the rater becomes lenient, whereas negative $Z_{SAI_{Difference}}$ values are interpreted as showing severity over time (Börkan, 2017).

Another index used to examine the change of peer raters' score over time is the term interaction (Wolfe et al., 2007). Time is added to the model as a dummy variable for the interaction index, and the interactions between the rater and the time variable are examined (Börkan, 2017).

Since the model data fit needs to be achieved in order to produce consistent and unbiased estimations in the multi-facet Rasch model, standardized residuals were examined. 1% of the standardized residual values were in the range of ±3, and 5% were in the range of ±2, indicating a sufficient model-data fit (Linacre, 2017). In the current study, 3 (0.03%) of the standardized residual values were found to be in the range of ±3, and 37 (3.70%) of them were found to be in the range of ±2 (total number of observations 25 raters x 8 groups x 4 criteria = 800). These results indicated that the model-data fit was acceptable, and the estimations were unbiased and consistent.

Results

Within the scope of the present study, the change in the scores of the peer raters over time was examined first. As the first day is determined as a baseline, the variation of the other days from the first day was examined. The logit values and standard differences obtained from the four measurements are given in Table 1.

Table 1

The Change of Peer Raters' Ratings Over Time

	Day 1		Day 2		Day 3		Day 4		$Z_{SAIdifference}$		
Rater	**Logit**	**SH**	**Logit**	**SH**	**Logit**	**SH**	**Logit**	**SH**	**2-1**	**3-1**	**4-1**
1	-1.10	0.65	-0.59	0.70	0.20	0.68	-0.12	0.63	0.53	1.38	1.08
2	-0.68	0.65	-0.59	0.70	-0.26	0.68	0.77	0.72	0.09	0.45	1.49
3	-0.25	0.65	-1.06	0.67	-0.26	0.68	-0.50	0.61	-0.87	-0.01	-0.28
4	0.17	0.65	-1.06	0.67	0.20	0.68	-0.50	0.61	-1.32	0.03	-0.75
5	1.57	0.74	1.59	0.77	0.67	0.68	0.77	0.72	0.02	-0.90	-0.77
6	1.57	0.74	0.46	0.74	1.15	0.70	0.77	0.72	-1.06	-0.41	-0.77
7	1.06	0.69	-0.08	0.73	0.67	0.68	-0.12	0.63	-1.13	-0.40	-1.26
8	1.57	0.74	-0.08	0.73	1.15	0.70	0.77	0.72	-1.59	-0.41	-0.77
9	1.57	0.74	-0.08	0.73	-2.60	0.68	-2.29	0.61	-1.59	**-4.15**	**-4.02**
10	-0.25	0.65	-0.59	0.70	-0.26	0.68	0.29	0.66	-0.36	-0.01	0.58
11	-1.10	0.65	-0.59	0.70	0.20	0.68	-0.12	0.63	0.53	1.38	1.08
12	-1.52	0.65	-0.59	0.70	-0.26	0.68	0.29	0.66	0.97	1.34	1.95
13	-0.68	0.65	-0.59	0.70	0.20	0.68	-0.12	0.63	0.09	0.94	0.62
14	0.17	0.65	-1.06	0.67	-0.73	0.69	2.24	1.09	-1.32	-0.95	1.63
15	-1.52	0.65	-0.59	0.70	-0.26	0.68	1.36	0.83	0.97	1.34	**2.73**
16	-1.52	0.65	-0.59	0.70	0.20	0.68	2.24	1.09	0.97	1.83	**2.96**
17	-0.68	0.65	3.15	1.09	-1.67	0.69	-0.87	0.60	**3.02**	-1.04	-0.21
18	-0.25	0.65	-0.59	0.70	1.15	0.70	-1.22	0.60	-0.36	1.47	-1.10
19	-1.52	0.65	-0.59	0.70	-0.26	0.68	-0.87	0.60	0.97	1.34	0.73
20	-0.25	0.65	1.01	0.75	-1.20	0.69	-0.87	0.60	1.27	-1.00	-0.70
21	1.57	0.74	0.46	0.74	1.15	0.70	-1.93	0.60	-1.06	-0.41	**-3.67**
22	1.57	0.74	1.59	0.77	0.67	0.68	0.29	0.66	0.02	-0.90	-1.29
23	0.17	0.65	-1.06	0.67	0.67	0.68	0.77	0.72	-1.32	0.53	0.62
24	-0.25	0.65	0.46	0.74	-0.26	0.68	-0.87	0.60	0.72	-0.01	-0.70
25	0.61	0.66	1.59	0.77	-0.26	0.68	-0.12	0.63	0.97	-0.92	-0.80
Ort.	0.00		0.00		0.00		0.00		-0.03	0.02	-0.07
SD	1.11		1.06		0.90		1.10		1.14	1.28	1.67

Those with thick font sizes represent raters who performed rater drift statistically.

When Table 1 was examined, it was seen that the scores of 25 peer raters from day 1 to day 2 decreased by -0.03 points on average. On the third day, it was seen that the scores increased by 0.02 on average and decreased by 0.07 on the last day. It was stated that for a significant group-level rater severity or leniency, $Z_{SAI_{Difference}}$ should be 0.50 and above between two-time measures (Swaminathan & Rogers, 1990). When the average point of the other three-time measurements compared to the baseline time was examined, it was found to be close to zero, and all times had almost the same level of severity or leniency. Although there was no significant rater drift at the group level, it was revealed that some raters at the individual level tended to be more severe or lenient over time. For example, on day 2, rater 17 displayed a more lenient behavior than on the first day. As a result, it was presented that a few raters at the individual level made more severe or lenient ratings over time, but there was no significant rater drift at the group level.

In addition to using standard differences in determining rater drift, rater drift was examined with the interaction term by examining common effects. Accordingly, within the scope of the present study, the time variable was included in the model, the rater x time interactions were examined, and the findings are given in Table 2.

Table 2

Rater X Time Interactions

	Day 1		**Day 2**		**Day 3**		**Day 4**		$I_{difference}$		
Rater	**Bias**	**SH**	**Bias**	**SH**	**Bias**	**SH**	**Bias**	**SH**	**2-1**	**3-1**	**4-1**
1	-0.72	0.59	-0.09	0.60	0.37	0.61	0.48	0.63	0.75	1.28	1.39
2	-0.55	0.60	-0.28	0.60	-0.19	0.60	1.17	0.71	0.32	0.42	1.85
3	0.10	0.61	-0.36	0.60	0.09	0.60	0.18	0.61	-0.54	-0.01	0.09
4	0.28	0.62	-0.55	0.60	0.28	0.61	0.00	0.61	-0.96	0.00	-0.32
5	0.35	0.72	0.23	0.71	-0.56	0.63	0.04	0.71	-0.12	-0.95	-0.31
6	0.46	0.72	-0.54	0.63	-0.03	0.66	0.16	0.71	-1.05	-0.50	-0.30
7	0.52	0.67	-0.39	0.61	0.08	0.63	-0.19	0.63	-1.00	-0.48	-0.77
8	0.57	0.72	-0.81	0.61	0.08	0.66	0.27	0.71	-1.46	-0.50	-0.30
9	2.12	0.72	0.73	0.61	-1.32	0.59	-1.25	0.59	-1.47	**-3.70**	**-3.62**
10	-0.18	0.61	-0.28	0.60	-0.19	0.60	0.70	0.66	-0.12	-0.01	0.98
11	-0.72	0.59	-0.09	0.60	0.37	0.61	0.48	0.63	0.75	1.28	1.39
12	-0.98	0.59	0.00	0.60	0.09	0.60	0.98	0.66	1.16	1.27	2.21
13	-0.45	0.60	-0.19	0.60	0.28	0.61	0.38	0.63	0.31	0.85	0.95
14	0.00	0.62	-0.83	0.60	-0.74	0.60	2.40	1.07	-0.96	-0.86	1.94
15	-1.16	0.59	-0.19	0.60	-0.09	0.60	1.84	0.81	1.15	1.27	**2.99**
16	-1.35	0.59	-0.37	0.60	0.09	0.61	2.50	1.07	1.16	1.70	**3.15**
17	-0.45	0.60	2.89	1.07	-1.16	0.59	-0.37	0.60	**2.72**	-0.84	0.09
18	-0.09	0.61	-0.19	0.60	1.08	0.66	-0.74	0.60	-0.12	1.30	-0.76
19	-0.70	0.59	0.27	0.60	0.36	0.60	0.08	0.60	1.15	1.26	0.93
20	0.00	0.61	1.08	0.66	-0.72	0.59	-0.28	0.60	1.20	-0.85	-0.33
21	1.19	0.72	0.19	0.63	0.70	0.66	-1.82	0.59	-1.05	-0.50	**-3.23**
22	0.46	0.72	0.35	0.71	-0.45	0.63	-0.31	0.66	-0.11	-0.95	-0.79
23	-0.10	0.62	-0.93	0.60	0.28	0.63	0.88	0.71	-0.96	0.43	1.04
24	-0.09	0.61	0.57	0.63	-0.09	0.60	-0.37	0.60	0.75	0.00	-0.33
25	0.10	0.63	0.88	0.71	-0.67	0.60	-0.19	0.63	0.82	-0.89	-0.33
Ort.	-0.06		0.04		-0.08		0.28		0.09	0.00	0.31
SD	0.75		0.79		0.55		0.99		1.06	1.18	1.61

Fixed (all = 0) chi-square: 116.3 d.f.: 100 significance (probability): .130

Variance explained by the interaction (%): 11.24

When Table 2 was examined, it was explored that the rater x time interaction was not statistically significant at the group level ($\chi2$ (100) = 116.30, p > .05). At the individual level, it was revealed that the scores of some raters at different times showed a drift. With interaction analysis, it was found that the standard differences gave similar results. The fact that the rater drifts, which were statistically significant in standardized differences, were also significant as a result of the interaction analysis indicates that both techniques are powerful in detecting rater drifts. The graph of the t-values for rater time interactions is given in Figure 1.

Figure 1 shows that there are more values outside the ±1.96 range, especially in the fourth time measurement. This finding provides evidence that peer raters' scores may vary over time. There is a smaller change in the scoring times of the 3rd time compared to the other times. The ratings of each rater over time are given in Figure 2.

Figure 1

T-values for Rater x Time Interactions

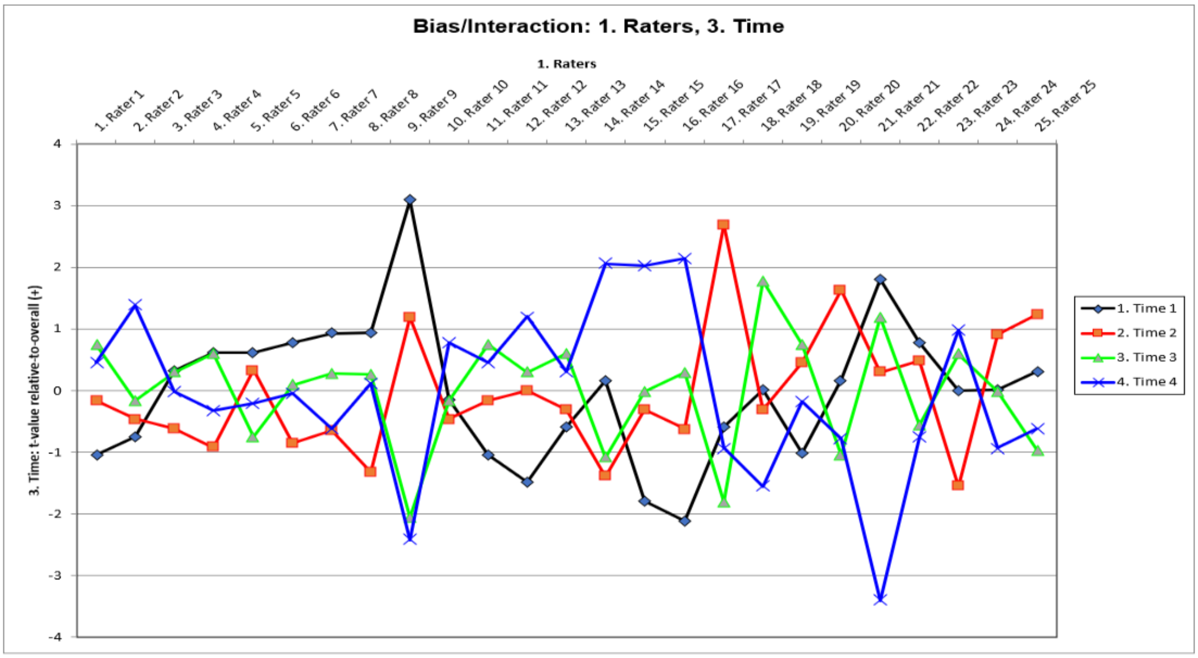
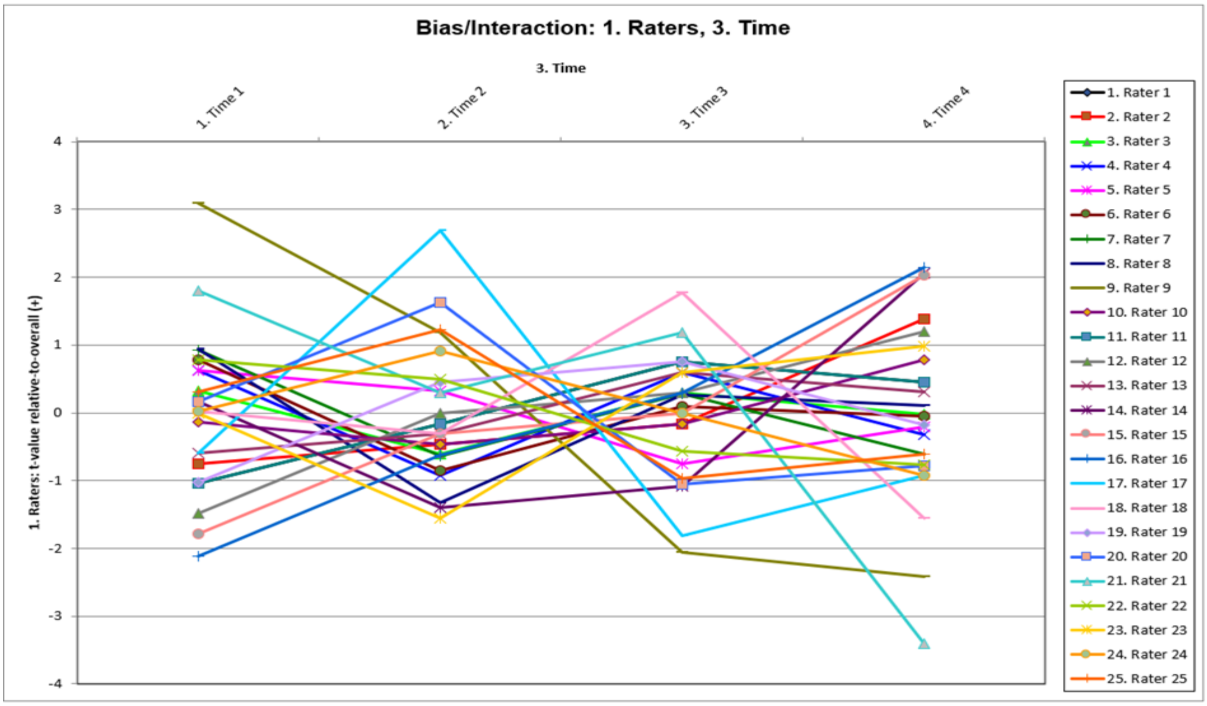


Figure 2

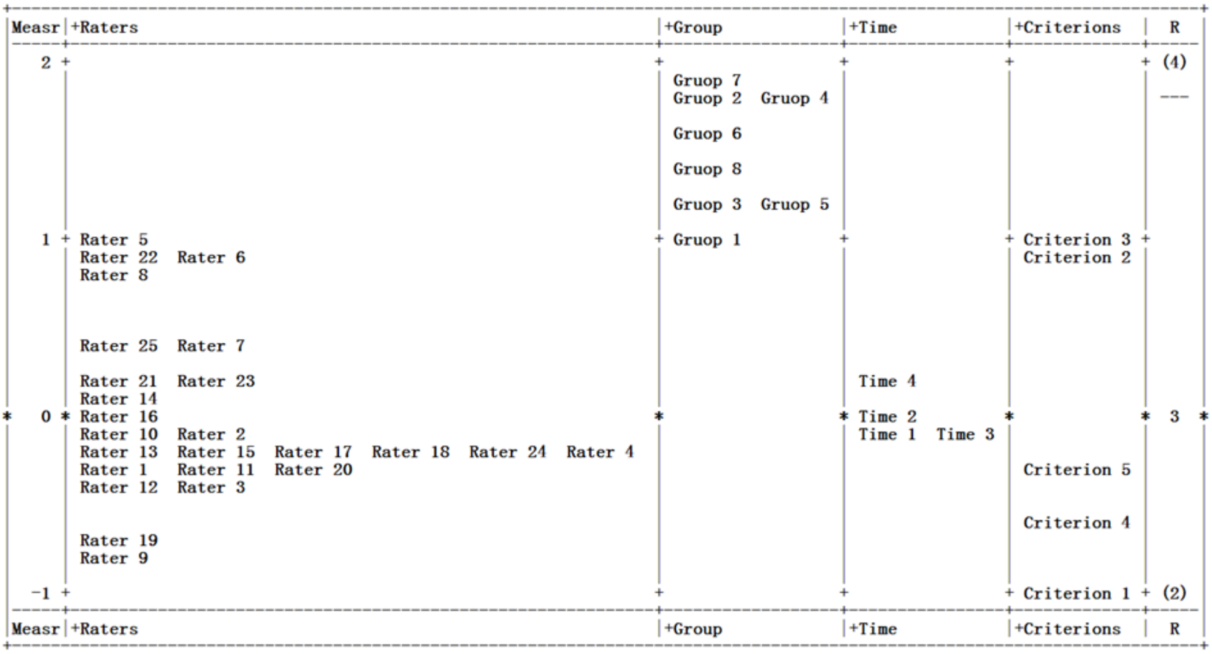
T-values for Each Raters' Ratings Over Time



When Figure 2 was examined, we observed that the t-values for raters 9, 17, and 21 were higher than the critical t-values, as well as rater drift over time. It was determined that the other raters had small rater drifts over time, but there were no statistically significant changes. Finally, the variable map, which allows examining all facets together, is given in Figure 3.

Figure 3

The Variable Map



```
+-------------------------------------------------------------------------------------------------------------------------+
|Measr|+Raters                                                     |+Group            |+Time          |+Criterions  |  R  |
|-----+------------------------------------------------------------+------------------+---------------+-------------+-----|
|   2 +                                                            +                  +               +             + (4) |
|     |                                                            | Gruop 7          |               |             |     |
|     |                                                            | Gruop 2  Gruop 4 |               |             | --- |
|     |                                                            |                  |               |             |     |
|     |                                                            | Gruop 6          |               |             |     |
|     |                                                            |                  |               |             |     |
|     |                                                            | Gruop 8          |               |             |     |
|     |                                                            |                  |               |             |     |
|     |                                                            | Gruop 3  Gruop 5 |               |             |     |
|     |                                                            |                  |               |             |     |
|   1 + Rater 5                                                    + Gruop 1          +               +  Criterion 3+     |
|     | Rater 22  Rater 6                                          |                  |               |  Criterion 2|     |
|     | Rater 8                                                    |                  |               |             |     |
|     |                                                            |                  |               |             |     |
|     |                                                            |                  |               |             |     |
|     |                                                            |                  |               |             |     |
|     | Rater 25  Rater 7                                          |                  |               |             |     |
|     |                                                            |                  |               |             |     |
|     | Rater 21  Rater 23                                         |                  | Time 4        |             |     |
|     | Rater 14                                                   |                  |               |             |     |
*   0 * Rater 16                                                   *                  * Time 2        *             *  3  *
|     | Rater 10  Rater 2                                          |                  | Time 1  Time 3|             |     |
|     | Rater 13  Rater 15  Rater 17  Rater 18  Rater 24  Rater 4  |                  |               |             |     |
|     | Rater 1   Rater 11  Rater 20                               |                  |               |  Criterion 5|     |
|     | Rater 12  Rater 3                                          |                  |               |             |     |
|     |                                                            |                  |               |             |     |
|     |                                                            |                  |               |  Criterion 4|     |
|     | Rater 19                                                   |                  |               |             |     |
|     | Rater 9                                                    |                  |               |             |     |
|  -1 +                                                            +                  +               +  Criterion 1+ (2) |
|-----+------------------------------------------------------------+------------------+---------------+-------------+-----|
|Measr|+Raters                                                     |+Group            |+Time          |+Criterions  |  R  |
+-------------------------------------------------------------------------------------------------------------------------+
```

When Figure 3 is examined, it is seen that the four facets within the scope of the research are placed on a single scale (logit scale). Thus, the four facets can be compared graphically with respect to each other. For example, who is the severity rater or who is the most successful group can be seen simultaneously.

Discussion and Conclusion

It is stated that peer assessment, which has increased in importance with alternative approaches, can enhance students' motivation in the learning environment (Topping, 2009) and contribute to their development of in-depth thinking and problem-solving skills (Patri, 2002). However, there is limited evidence for the validity of peer-reviewed scores. Therefore, it is considered necessary to carry out studies on the validity of the scores obtained from peer assessment, which is gaining importance so far.

In this study, rater severity drift, one of the rater effects, was examined in evaluating students' presentations that were spread through time by their peers. In the analyses made in peer assessment applications, it was tried to observe whether there was peer rater severity drift in the following three scoring days, compared with the first day, which was grounded based on the first day of scoring. In the study conducted with middle school 7th-grade students, peer assessment lasted four days, and peer raters used rubrics to evaluate the presentations. Accordingly, whether the peer raters drifted their peer severity at the individual or group level during the presentation was analyzed with the help of MFRM analysis. Standardized differences and interaction term (rater x time) approaches were used to determine rater drift at the group level. The results obtained from both approaches showed similar characteristics. According to the results, it was observed that the scores of the peer raters at the group level became severe from “Day 1” to “Day 2”, and their scores were more lenient on “Day 3” and became severe again on “Day 4”, yet these differences were not significant at the group level. This finding indicates that the students did not make a biased scoring while evaluating the group performance according to the groups. Although there was no significant rater drift at the group level, it was observed that some students at the individual level showed rater drifts over time. In determining rater drift at the individual level, the raters were compared among themselves using the standard differences and common effects (rater x time) approaches. The results obtained from both approaches showed similar characteristics. Accordingly, three of the students participating in the study demonstrated a rater drift over time. Two students (numbered 9 and 21) scored more severely over time, while one student (numbered 17) scored more leniently. Finally, when the variability map enabled us to examine all facets together, it was found that the most lenient rater was 5 and the most severe rater was 9. In the presentations, it was revealed that the most successful group was the 7th and the most unsuccessful group was the 1st group, that the presentations were consistently ranked by the raters according to their qualities, and it was determined that the drifts on other days were close to each other, except for the 4th day, in time measurements. According to the change map, it was also found out that the criterion that the groups had the least difficulty with was the 3rd criterion, "Material Use", and the criterion they had the most difficulty with was the 1st criterion, "Content".

The study group was an experienced one in the use of peer assessment approach and rubrics in the educational environment. The reason why the students did not make rater drift on a group basis may be that they included peer assessment practices in the learning process and therefore, they were experienced in this regard. It is stated in the related literature that rater education reduces rater effect (Hauenstein & McCusker, 2017). Another reason why students did not have rater drift at the group level may be that the presentation scores were spread over a short period of time (4 days). Harik et al. (2009) have stated that in studies whose scoring was done within days or weeks, rater drift could be at a minimum level when compared to studies whose scoring was spread over months or years.

Although there was no rater drift at the group level, it was observed that a small portion of the students (3 students) made more lenient ratings over time. Previous literature has different findings on this issue. While one study has shown that leniency increased over time (Lunz & Stahl, 1990), four studies have shown increasing severity (Congdon & McQueen, 2000; McLaughlin et al., 2009; Myford, 1991; Pinot de Moira et al., 2002), and two other studies have demonstrated positive and negative drift for a small proportion of their raters (Börkan, 2017; Myford & Wolfe, 2009).

Analyzing rater drift at the individual level, we found that two students displayed severity drift and three students displayed lenient drift. In the literature, it is seen that the leniency of raters, which differs in the peer assessment process, is quite common (Erman-Aslanoğlu et al., 2020; Farrokhi et al., 2012). Considering both standard differences and interaction/bias analysis, we can conclude that both methods can be used to detect the rater drift of the same individuals separately. Therefore, it will suffice to choose one method for future research. Moreover, considering that interaction analyses are systemic errors, they have a negative impact on the validity of measurements obtained from these individuals (Messick, 1996). As a result, evidence was provided for the reliability and validity of the measurements at both the group and individual levels in the peer assessment process, and it was determined that some peers had an effect on the validity of the measurements at the individual level. However, there was no statistically significant effect on the validity and reliability of the measurements at the group level. The exclusion of students with rater drift from the scoring will, therefore, contribute to the reliability and validity of the measurements if the evaluation of the students is crucial.

This research is limited to 7th-grade level and oral presentation skills. Researchers can investigate the effect of rater drift on peer ratings over time by conducting similar studies at different grades and with different skills. Researchers can also examine the effect of peer drift in terms of students who actively use the peer assessment approach in the teaching environment and who do not use this assessment approach. To reduce rater drifts that occur over time in peer assessment, rater training can be designed. Thus, it is expected that the measurements obtained in the performance assessment will contribute to the reliability and validity.

Declarations

Author Contribution: Aslıhan Erman Aslanoğlu: Theoretical framework, literature review, methodology, data collection, data analysis, discussion, writing the original draft and review & editing. Mehmet Şata: Theoretical framework, methodology, data analysis, discussion and review & editing.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: The study was approved by the Ufuk University Ethics Committee (Research code: 2021-49, dated 09.06.2021)

References

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching*, *53*(1), 27-31. https://doi.org/10.3200/CTCH.53.1.27-31

Azarnoosh, M. (2013). Peer assessment in an EFL context: attitudes and friendship bias. *Language Testing in Asia*, *3*(1), 1-10. https://doi.org/10.1186/2229-0443-3-11

Baird, J. A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability. A Comparative exploration from the perspectives of generalisability theory, Rash model and multilevel modelling*. Oxford University Centre for Educational Assessment. https://core.ac.uk/download/pdf/15171449.pdf

Batmaz, H., Türk, N., Kaya, A., & Yıldırım, M. (2022). Cyberbullying and cyber victimization: examining mediating roles of empathy and resilience. *Current Psychology*, 1-11. https://doi.org/10.1007/s12144-022-04134-3

Black, P., Harrison, C., & Lee, C. (2003). *Assessment for learning: Putting it into practice*. McGraw-Hill. http://www.mcgraw-hill.co.uk/html/0335212972.html

Börkan, B. (2017). Rater severity drift in peer assessment. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(4), 469-489. https://doi.org/10.21031/epod.328119

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

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178. https://psycnet.apa.org/doi/10.1111/j.1745-3984.2000.tb01081.x

Dochy, F. (2001). A new assessment era: different needs, new challenges. *Learning and Instruction, 10*(1), 11-20. https://doi.org/10.1016/S0959-4752(00)00022-0

Donnon, T., McIlwrick, J., & Woloschuk, W. (2013). Investigating the reliability and validity of self and peer assessment to measure medical students' professional competencies. *Creative Education*, *4*(6A), 23-28. http://dx.doi.org/10.4236/ce.2013.46A005

Erman Aslanoğlu, A., Karakaya, İ., & Şata, M. (2020). Evaluation of university students' rating behaviors in self and peer rating process via many facet rasch model. *Eurasian Journal of Educational Research*, 20(89), 25-46. https://dergipark.org.tr/en/pub/ejer/issue/57497/815802

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287–322. https://doi.org/10.2307/1170785

Farrokhi, F., Esfandiari, R. ve Dalili, M.V. (2011). Applying the Many-Facet Rasch Model to detect centrality in self-Assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal 15 (Innovation and Pedagogy for Lifelong Learning)*, 70-77. http://www.idosi.org/wasj/wasj15(IPLL)11/12.pdf

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

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education*, *25*(12), 1509–1528. https://doi.org/10.1080/0950069022000038268

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. 2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*(1), 43-58. https://doi.org/10.1111/j.1745-3984.2009.01068.x

Hauenstein, N. M. A. & McCusker, M. E**.** (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment, 25*, 253-266. https://psycnet.apa.org/doi/10.1111/j.1745-3984.2009.01068.x

Humphris GM, & Kaney S. (2001). Examiner fatigue in communication skills objective structured clinical examinations. *Medical Education*, *35*(5), 444-449. https://doi.org/10.1046/j.1365-2923.2001.00893.x

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

Kaya, A., Türk, N., Batmaz, H., & Griffiths, M. D. (2023). Online gaming addiction and basic psychological needs among adolescents: the mediating roles of meaning in life and responsibility. *International Journal of Mental Health and Addiction*, 1-25. https://doi.org/10.1007/s11469-022-00994-9

Kassim, A.N.L. (2007, June 14-16). *Exploring rater judging behaviour using the many-facet Rasch model* [Conference Session]. The Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2), Universiti Utara, Malaysia. http://repo.uum.edu.my/id/eprint/3212/

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

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

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement: Classroom application and practice* (10th ed.). John Wiley & Sons. https://l24.im/jV6yYCJ

Kutlu, Ö., Doğan, C. D., & Karaya, İ. (2014). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme* [Determining student success: Assessment based on performance and portfolio]. Pegem. https://l24.im/k5cn

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399-418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

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

Liu, H., & Brantmeier, C. (2019). "I know English": Self-assessment of foreign language reading and writing abilities among young Chinese learners of English. *System, 80*, 60-72. https://doi.org/10.1016/j.system.2018.10.013

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions, 13*(4), 425 444. https://psycnet.apa.org/doi/10.1177/016327879001300405

McLaughlin K, Ainslie M, Coderre S, Wright B & Violato C. (2009). The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. *Medical Education, 43*(10), 989–992. https://doi.org/10.1111/j.1365-2923.2009.03438.x

Messick, S. (1994). Alternative modes of assessment, uniform standards of validity. *ETS Research Report Series, 2*, 1-22. https://doi.org/10.1002/j.2333-8504.1994.tb01634.x

Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). National Center for Education Statistics.

Moore, B. B. (2009). *A consideration of rater effects and rater design via signal detection theory* (Publication No. 3373803). [Doctoral dissertation, Columbia University]. ProQuest Dissertations & Theses Global.

Mulqueen C., Baker D. & Dismukes R.K. (2000, April). *Using multifacet rasch analysis to examine the effectiveness of rater training* [Conference Session]. 15th Annual Meeting of the Society for Industrial/Organizational Psychology, New Orleans, LA. https://www.air.org/sites/default/files/2021-06/multifacet_rasch_0.pdf

Myford, C. M. (1991, April 3-7). *Judging acting ability: The transition from notice to expert* [Conference Session]. Annual Meetin of the American Educational Research Association, Chicago IL. https://files.eric.ed.gov/fulltext/ED333032.pdf

Myford, C. M., & Wolfe, E. M. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Aplied Measurement*, *4*(4), 386-422. https://psycnet.apa.org/record/2003-09517-007

Myford, C. M., & Wolfe, E. M. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. *Journal of Educational Measurement*, *46*(4), 371-389. https://psycnet.apa.org/doi/10.1111/j.1745-3984.2009.00088.x

Oosterhof, A. (2003). *Developing and using classroom assessment* (3th ed.). Merrill/Prentice Hall. https://l24.im/OCKvkg2

Orsmond P, Merry S, Reiling K (2000) The use of student-derived marking criteria in peer and self-assessment. *Assessment &Evaluation Higher Education*, *25*(1), 21–38. https://doi.org/10.1080/02602930050025006

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

Pinot de Moira, A., Massey, C., Baird, J., & Morrissey, M. (2002). Marking consistency over time. *Research in Education*, *67*(1), 79–87. https://doi.org/10.7227/RIE.67.8

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207. https://psycnet.apa.org/doi/10.1177/014662169001400208

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370. https://psycnet.apa.org/doi/10.1111/j.1745-3984.1990.tb00754.x

Şata. M. (2020a). Quantitative research approaches [Nicel araştırma yaklaşımları]. In Oğuz. E. (Ed.). *Research methods in education [Eğitimde araştırma yöntemleri*] (p. 77-98). Eğiten Kitap publishing.

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. S. Segers, Dochy,R., & E. C. Cascallar (Eds.), *In optimising new modes of assessment: In search of qualities and standards* (pp. 55-87). Springer Dordrecht. https://doi.org/10.1007/0-306-48125-1

Topping, K. (2009). Peer assessment. *Theory Into Practice*, *48*(1), 20-27. https://doi.org/10.1080/00405840802577569

Topping, K. (2017). Peer assessment: learning by judging and discussing the work of other learners. *Interdisciplinary Education and Psychology, 1*(1), 1-17. https://doi.org/10.31532/InterdiscipEducPsychol.1.1.007

Wolfe, E. W., Myford, C. M., Engelhard Jr. G., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP® English literature and composition examination using benchmark essays.* College Board. https://files.eric.ed.gov/fulltext/ED561038.pdf

Yurdabakan, İ., & Cihanoğlu, M. O. (2009). The effects of cooperative reading and composition technique with the applications of self and peer assessment on the levels of achivement, attitude and strategy use. *Dokuz Eylul University The Journal of Graduate School of Social Sciences*, *11*(4), 105-123. https://l24.im/VbHg

Appendix

Data Collection Tool

Criteria	1	2	3	4
Content	Topic is irrelevant and focused; presentation contains multiple fact errors.	Topic should be more focused and relevant; presentation contains some fact errors or omissions.	Topic is adequately focused and relevant; major facts are accurate and generally complete.	Topic is tightly focused and relevant; presentation contains accurate information with no fact errors.
Coherence	Ideas are not presented in proper order; transition is lacking between major ideas; several parts of the presentation are wordy or unclear.	Some ideas are not presented in proper order; transition markers are needed between some ideas; some parts of the presentation are wordy or unclear.	Most ideas are in logical order with adequate transitions between most major ideas; presentation is generally clear and understandable.	Ideas are presented in logical order with effective transitions between major ideas; presentation is clear and concise.
Use of Material	No material is used in the presentation.	Presentation is supported with a relevant material.	Presentation is supported with 2 different relevant materials.	Presentation is supported with 3 different relevant materials.
Communication	Inadequate voicing or energy, too slow or too fast pacing, poor pronunciation, distracting gestures or posture, unprofessional appearance, and visual aids poorly are used.	Neither adequate nor inadequate voicing and energy; slow or fast pacing; some distracting gestures or posture; adequate appearance; few visual aids are used.	Adequate voicing and energy; generally good pacing and intonation; few or no distracting gestures; professional appearance; adequate visual aids are used.	Proper voicing and energy; good pacing and intonation; no distracting gestures; professional appearance; effective and adequate visual aids are used.
Use of Time	The presentation exceeds or lags behind the time limit.	The presentation does not comply with the time limit (+/- 3).	The presentation does not comply with the time limit (+/- 2).	The presentation is completed on time.