




Examining Rater Biases of Peer Assessors in Different Assessment Environments¹

Sabahattin YEŞİLÇINAR², Mehmet ŞATA³

¹Muş Alparslan University- Department of English Language Teaching, Muş, Turkey  0000-0001-6457-0211

³Ağrı İbrahim Çeçen University- Department of Measurement and Evaluation in Education, Ağrı, Turkey  0000-0003-2683-4997

ARTICLE INFO

Article History

Received 07.06.2021

Received in revised form
17.08.2021

Accepted 22.08.2021

Article Type: Research
Article

ABSTRACT

The current study employed many-facet Rasch measurement (MFRM) to explain the rater bias patterns of EFL student teachers (hereafter students) when they rate the teaching performance of their peers in three assessment environments: online, face-to-face, and anonymous. Twenty-four students and two instructors rated 72 micro-teachings performed by senior Turkish students. The performance was assessed using a five-category analytic rubric developed by the researchers (Lesson Presentation, Classroom Management, Communication, Material, and Instructional Feedback). MFRM revealed the severity and leniency biases in all three assessment environments at the group and individual levels, drawing attention to the less occurrence of biases anonymous assessment. The central tendency and halo effects were observed only at the individual level in all three assessment environments, and these errors were similar to each other. Semi-structured interviews with peer raters (n = 24) documented their perspectives about how the anonymous assessment affected the severity, leniency, central tendency, and halo effects. Besides, the findings displayed that hiding the identity of the peers develops the reliability and validity of the measurements performed during peer assessment.

© 2021 IJPES. All rights reserved

Keywords:

Many-facet Rasch analysis; rater bias; anonymous assessment; peer assessment; teacher assessment.

1. Introduction

Peer assessment requires individuals' active participation in the assessment process and helps them play both assessors and assesses (Li, 2017). Peer assessors usually give feedback to the product of others for either summative grading or formative feedback, or an integration of both (Li & Gao, 2015). When students see their peers work from the assessor's viewpoint, they become more professional assessors, determining the pros and cons of other students' works based on a rich set of items for assessment (Cho & MacArthur, 2010). However, when acting as assesses, students consider and reflect upon peer feedback and develop their work (Li et al., 2012). Student's ability to perform learning and assessment tasks rests on how they perceive these activities (Boud & Soler, 2016). If they consider themselves as talented peer assessors, their engagement in peer assessment will increase, and they will believe in the usefulness of peer assessment (Vanderhoven et al., 2015). The more students experience the peer assessment process, the more likely they will make sound peer assessment judgements (Panadero, 2016). Thanks to the latest change in the teacher training program (May 30, 2018), Turkey has started to consider using feedback in teacher education. Although using peer feedback is not at the desired level, it is used while assessing oral presentation skills, in particular for students' teaching

¹The preliminary findings of this study were presented in International Pegem Conference on Education (IPCEDU-2020).

²Corresponding author's address: Muş Alparslan University- Department of English Language Teaching, Muş, Turkey
e-mail: sycinar@yahoo.com

Citation: Yeşilçınar, S. & Şata, M. (2021). Examining rater biases of peer assessors in different assessment environments. *International Journal of Psychology and Educational Studies*, 8(4), 136-151. <https://dx.doi.org/10.52380/ijpes.2021.8.4.622>

performance within the scope of teaching practice course (Güneş & Kılıç, 2016). Therefore, this study is significant because it examines the impact of various assessment environments on rating quality.

1.1. Anonymity in Peer Assessment

A conceptual basis regarding the possible effect of anonymity indicates that students' feedback will differ depending on whether their identities as assessors are revealed (Yu & Sung, 2016). Publicly assessing their peers may cause students to feel uncomfortable and experience stress (Pope, 2005). Therefore, anonymity can be suggested to help alleviate the interpersonal burden on students (Yu & Liu, 2009), avoid the pressure of friendships (Cheng & Tsai, 2012), foster higher participation (Chester & Gwynne, 2006; Vickerman, 2009), and to provide a feeling of psychological safety (Miyazoe & Anderson, 2011; Yu & Liu, 2009), referring to "a shared belief denoting one's emotional ability to take an interpersonal risk without fearing negative consequences about one's well-being, self-image, and status" (Kahn, 1990; Zhang et al., 2010 as cited in Rotsaert et al., 2018, p. 78). Emphasising the crucial role of anonymity in decreasing reciprocity effects, Freeman and McKenzie (2000) advocate that anonymity for the assessor may lead to fairer assessment. Additionally, anonymity for assessors diminishes the fear of disapproval and help assessors experience less peer pressure when giving low scores or negative feedback (Vanderhoven et al., 2015). Furthermore, there is over-scoring if assessors and assesseees are close friends (Panadero et al., 2013) because 'Relationships between students can colour opinions' (Papinczak et al., 2007, p. 180). These liberating effects are the basis of preferring anonymity to environments in which students are familiar with each other (e.g., Hosack 2004).

1.2. Objectivity and Rater Bias in Performance Assessment

Performance assessment refers to observing individuals' actions, determining their strengths and deficiencies based on the observation, developing their strengths, and determining and overcoming their shortcomings (Bennett, 1998). Performance assessment differs from traditional assessment. For example, performance assessment is based on real-life sections, focuses more on processes rather than the product, determines an individual's strength and weakness, and urges the individual to think more and solve problems (Brown & Hudson, 1998; Khaatri et al., 1995; Moore, 2009). The crucial concern about performance assessment is the objectivity of scoring an individual's performance as it is not easy to assess performance objectively, unlike traditional assessments (e.g. fixed response assessment) (Romagnano, 2001). Factors reducing objectivity in the performance assessment process are defined as rater bias or effect (Farrokhi et al., 2011; Cetin & Ilhan, 2017).

Kingsbury (1922) classifies rater biases as severity, central tendency, and halo effects. Raters can demonstrate serious differences in the severity and leniency of their rating due to their subjectivity or inconsistency (e.g., Bonk & Ockey, 2003; Weigle, 1998). If one or more rater biases occur when assessing performance, the bias amount of predictions will be high. That is, the predictions will cause unreliable measurements. Rater bias is attributed to construct-irrelevant variance, posing a direct threat of validity (Farrokhi et al., 2011; Messick, 1996). Therefore, how raters introduce construct-irrelevant variance is under discussion (Trace et al., 2017). Central tendency occurs when a rater shows more tendency towards the middle category than other categories (Royal & Hecker, 2016). According to Engelhard (1994), while some raters overuse extreme categories, others prefer the middle category. Anastasi (1976) states that avoiding using extreme scores decreases both reliability and validity since it significantly reduces the variability of the scoring. Halo effect refers to the tendency of any rater to provide the same scores to different individual characteristics when assessing the individual's performance (MyFord & Wolfe, 2004).

1.3. Many-Facet Rasch Measurement

Considering all sources of variability, MFRM also focuses on the interactions of these sources of variability (Abu Kassim, 2007). Unlike traditional two-category measurements, MFRM is defined as an extension of the partial credit model developed by Master (1982), which makes it possible to evaluate multi-category measurement tools (Myford, 2002). Being a linear model, MFRM calibrates all parameters and enables observed ratings to be transformed into a logit scale (Bond & Fox, 2015). Thanks to the logistic transformation of the log odds ratio, independent variables are seen as dependent variables (Esfandiari, 2015). In MFRM, each source of variability affecting the performance scores of individuals is called a facet (Sudweeks, Reeve & Bradshaw, 2005).

Regarding studies examining performance-based peer assessments, MFRM is proposed to give evidence of the reliability and validity of the scores and to eliminate the limitations of classical approaches (Baird et al., 2013; Kim et al., 2012). The factors that may affect the scoring during the performance assessment process are not limited to the individuals' ability or the difficulty levels of the items; rater-related issues may also cause variability in the scores of learner's performance (Baird et al., 2013). Therefore, MFRM becomes an appropriate option in the performance assessment where rater bias is effective. MFRM is also considered as a psychometrically stronger model than the classical test theory as it can determine the interactions between the facets (different error sources) (Haiyang, 2010), take into account multiple error sources at the same time and generate higher ability predictions for validity (Ilhan, 2016), and provide information on each facet at both group and individual levels (Barkaoui, 2013).

The current study investigated the rater biases that affected the students during peer assessment in terms of the assessment environments. Additionally, the development and assessment status of the students' performance was examined over time through intervals. Accordingly, the following questions were asked:

- Considering the assessment of students' teaching performance, do teacher and peer assessment differ according to the assessment environments?
- In terms of the assessment environments, (i) Do the severity and leniency biases of teachers and peer raters differ? (ii) Does the central tendency effect of teachers and peer raters differ? (iii) Does the halo effect of peer raters differ?

2. Methodology

2.1. Research Design

This paper adopted the explanatory sequential design (Schoonenboom & Johnson, 2017) to determine the rater biases occurring when students assess their peers' teaching performance. After collecting and analysing quantitative data, the researchers used the qualitative part to explain the initial quantitative results. Thus, quantitative analyses were carried out to identify the rater bias, and then qualitative data analyses were performed to determine the reason(s) for rater biases.

2.2. Research Sample

Participants were 24 senior students (17 females and 7 males) and two academicians (one female and one male) of English Language Teaching in a state university. The age average was 41 for academicians and 22 for students. Academics were chosen to see whether students' assessments in intervals reflect students' micro-teaching performance. The students were considered to have similar levels for two reasons. First, they matriculated at the same university, indicating that their university entrance scores were very close. Second, both the instructors' opinions and the micro-teaching scores of the first semester were taken into account. Those whose scores were between 70 and 85 were requested to participate in the study. There were 34 volunteers. After the corresponding researcher arranged a meeting and informed them about the process, only 24 students remained. After they signed a constant form, they were randomly recruited to the groups. The participants' names were changed to maintain anonymity (see Table 1). Every week, each group had two micro-teachings on different days.

Table 1. Participants

Assessment Environment	Participants
Face-to-face	FP1, FP2, FP3, FP4, FP5, FP6, FP7, FP8
Online	OP1, OP2, OP3, OP4, OP5, OP6, OP7, OP8
Anonymous	AP1, AP2, AP3, AP4, AP5, AP6, AP7, AP8

In the face-to-face assessment, peers were assessed and given feedback immediately after the micro-teaching. However, in anonymous assessment, peers were asked to assess and give feedback through Edmodo (<https://www.edmodo.com/>) on the same day. Participants in the anonymous group used different nicknames for each week to hide their identities. To find whether the difference between face-to-face and anonymous assessments is due to anonymity (not due to other variables), the researchers created a third assessment environment (online). The online and anonymous assessments process was the same; the only difference was that assessors were known in the online environment. Finally, the corresponding researcher interviewed all

participants separately to investigate their views about the assessment environments. Besides, they talked with participants in the anonymous environment to learn students' nicknames to examine a participant's assessments and feedback.

2.3. Data Collection

2.3.1. Quantitative data tool

Table 2 below presents the content validity ratios for the basic and sub-items included in the draft form.

Table 2. CVR Values Related to Items Included in the Draft Form of the Measurement Tool

Basic items	Sub-items	Necessary	Should be corrected	Unnecessary	CVR
Lesson Presentation	Starts a lesson in an engaging way.	7	0	0	1.000
	Uses time efficiently.	7	0	0	1.000
	Uses various teaching methods and techniques appropriately.	6	1	0	0.714
	Completes the whole course.	7	0	0	1.000
	Provides relevant examples and demonstrations to illustrate concepts and skills.	7	0	0	1.000
	Assigns tasks appropriate to student level.	7	0	0	1.000
	Facilitates smooth and effective transitions between instructional activities.	6	1	0	0.714
	Summarises the main point(s) at the end of the lesson or instructional activities.	7	0	0	1.000
Classroom Management	Manages discipline problems	7	0	0	1.000
	Creates a stimulating and effective environment for learning	7	0	0	1.000
	Creates opportunities for and manage individual, partner, group, and whole class work.	7	0	0	1.000
Communication	Communicates effectively with students	6	1	0	0.714
	Gives clear explanations and instructions	5	0	2	0.429*
	Speaks fluently and precisely	7	0	0	1.000
	Decides when it is appropriate to use the target language and when not to.	7	0	0	1.000
Material	Prepares appropriate tools and materials	6	0	1	0.714
	Uses appropriate tools and materials	4	1	2	0.143*
	Uses material in an organised manner.	7	0	0	1.000
	Uses material at an appropriate pace.	7	0	0	1.000
Instructional Feedback	Provides prompt feedback on assigned work	5	1	1	0.429*
	Provides sustaining feedback after an incorrect response	7	0	0	1.000
	Uses appropriate type and amount of feedback for target behaviours	6	0	1	0.714
Content Validity Index (CVI)					0.925

*CVR < .622

To assess students' teaching performance, researchers developed the *Analytic rubric for performance* (ARP) based on the literature, following a systematic process that includes certain stages (Akpınar, 2019). This is because preparing rubrics without considering this process negatively affects the validity and reliability of the assessment (Moskal, 2000). While developing the analytical rubric, the recommendations of Goodrich (1997), Haladyna (1997), Kutlu et al. (2014), and Moskal (2000) were considered. In this context, a systematic process has been followed. First, the purpose of the measurement tool was determined, and measuring student teachers' teaching performance was considered the main objective. Then, the literature was searched to determine the criteria (Newby et al., 2007; Schools & Chesterfield, 2015). After determining the criteria, two academicians who were experts in measuring the relevant performance prepared the draft version of the

measurement tool. A draft form was sent to seven field experts to collect evidence for content validity. Then, two field experts and a measurement and evaluation specialist were consulted to determine how many levels each criterion should have. It was concluded that the quadruple rating would be appropriate in measuring the relevant structure. Finally, regarding the type of rubric, it was decided that the most appropriate type for the relevant performance would be analytical. The researchers first prepared a draft form consisting of 22 items and then sent it to seven experts. They were asked to assess the items through a measurement tool with triple rating: (1) necessary, (2) should be corrected, and (3) unnecessary. The Lawshe (1975) approach was taken into account for the content validity of the items in the rubric. Since seven experts were administered, the minimum content validity ratio (CVR) was accepted as 0.622 ($p = 0.05$) to acknowledge that an item measures the relevant structure (Wilson et al., 2012). Accordingly, the items whose CVR values were equal to or higher than 0.622 were included in the main form of the measurement tool. Three sub-items were lower than the threshold value of the CVR. Therefore, they were removed from the rubric, and the final form consisted of 19 sub-items and five basic items. After calculating the CVR of the items, the content validity index (CVI) was obtained for the whole measurement tool (Lawshe, 1975). CVI is the construct validity process (Lawshe, 1985) and was 0.925, indicating a higher value. This shows that the related instrument could measure students' teaching performance. In line with the literature and the opinions of the field experts, the measurement tool adopted four categories (1 = poor, 2 = fair, 3 = good, and 4 = outstanding).

Considering construct validity, the tetracolic exploratory factor analysis (EFA) was carried out. During the tetracolic EFA, each student's scores in the first interval were taken into consideration. Each item was measured 216 (24×9) times, as seven peers and two teachers assessed each student. The tetracolic EFA was made on this data set using the Mplus package program. As a result, 19 items were found to be collected under a single latent factor. The instrument was a four-point scale, so handling it at the ordinal scale level rather than the continuous scale will contribute to the validity and reliability of the measurements. Thus, the researchers used tetracolic EFA, which provides more consistent estimates for ordinal scales.

Table 3. Model-Data Fit Indices for the Tetracolic EFA

Model-Data Fit Criteria	Acceptable Fit	Estimates
χ^2/sd value	$2 \leq \chi^2/sd \leq 5$	2.595
RMSEA (%90GA)	$0.05 \leq RMSEA \leq 0.10$	0.086 (0.076-0.096)
CFI	$0.90 \leq CFI < 0.95$	0.994
TLI (NNFI)	$0.90 \leq NNFI < 0.95$	0.993
SRMR	$0.05 \leq SRMR < 0.10$	0.041

The model data fit criteria for the tetracolic EFA were at acceptable values. Cronbach α reliability coefficient was used for the reliability of the measurements and found to be 0.977 (95% Confidence Interval: 0.972-0.981). Findings indicated that the measurements of the ARP gave valid and reliable results.

2.3.2. Qualitative data tool

Semi-structured interviews were conducted to "help explain, or elaborate on, the quantitative results obtained in the first phase" (Ivankova et al., p. 5). They were conducted in students' native language to ensure the quality and quantity of the data (Mackay & Gass, 2005). Before data collection, the interview protocol was reviewed for accuracy and then piloted. All interviews were recorded and took place mostly in the corresponding researcher's office. Each lasted between 35 and 45 minutes. The total duration was 945 minutes.

2.4. Data Analysis

2.4.1. Quantitative data analysis

MFRM, Mann Whitney U test, and the Friedman test were used for data analysis. There were five facets (individual, raters, assessment items, rater type, and interval). A completely crossed design was used since students were scored by all raters and from all items. MFRM was applied under this pattern. Some assumptions must be met for the analyses made using MFRM to make consistent estimates. These assumptions are unidimensionality, and local independence, model-data fit. To use MFRM, the instrument needs to measure a single construct. As all items in the ARP measured a single construct, the first assumption (unidimensionality) was met. Then, the local independence assumption was tested using the G2 statistics developed by Chen and Thissen (1997). The local independence requirement was determined to be met for

each item because the standardised LD χ^2 values estimated between each variable pair were below 10, and the marginal fit χ^2 indices estimated for each item were close to zero. Considering model-data fit, the number of standardised residual values out of the ± 2 range should not exceed 5% of the total number of observations and standardised residual values out of the ± 3 range should not exceed 1% of the total data (Linacre, 2017). Thus, the model-data fit was assumed to be achieved for three assessment types since the total number of observations for anonymous assessment was $8 \times 9 \times 19 \times 3 = 4.104$, while the number of standardised residual values out of the ± 2 range was 258 (6.29%), and the number of standardised residual values out of the ± 3 range was 72 (1.75%). The total number of observations for face-to-face assessment was $8 \times 9 \times 19 \times 3 = 4.104$, while the number of standardised residual values out of the ± 2 range was 227 (5.53%), and the number of standardised residual values out of the ± 3 range was 7 (0.50%). The total number of observations for online assessment is $8 \times 9 \times 19 \times 3 = 4.104$, while the number of standardised residual values out of the ± 2 range was 270 (6.58%), and the number of standardised residual values out of the ± 3 range was 82 (1.99%). As a result, all assumptions were met, and thus analyses were performed.

2.4.2. Qualitative data analysis

Nvivo was used to analyse the qualitative data. However, the qualitative data were initially manually analysed as a piloting process (Welsh, 2002) so that the researcher could observe, control, and manage the data. This process enabled the researchers to take precautions for data loss. Content analysis was used, which means that all categories were determined while analysing the transcripts. First, all interviews were transcribed. The researchers adopted an 'edited transcription' (Hansen, 2003, p. 136) to ease the analysis process and read the data for getting a general idea. Then, they examined and categorised the data independently.

3. Findings

This study examined whether there is a significant difference between peer and teacher assessment in assessment environments (anonymous, face-to-face, and online) over time (three intervals/measurements). Besides, rater biases (severity, leniency, halo, central tendency, differential severity, and differential leniency) were studied considering the assessment environments. A model was established for each assessment environment, and three Rasch analyses were performed. Findings were presented under the relevant headings.

3.1. Teacher and Peer Assessment in Terms of Assessment Environment

The Mann Whitney U test was used to determine whether teacher and peer assessment differ statistically according to the assessment environments and assessment intervals.

Table 4. Mann Whitney U Test on Teacher and Peer Assessment

Assessment environments	Interval	Assessment Types	N	Mean Rank	Sum of Rank	U	Z	p	η^2
Face-to-face	1. Measure	Peer	56	43.22	2420.50	71.50	-5.10	0.000*	0.60
		Teacher	16	12.97	207.50				
	2. Measure	Peer	56	43.98	2463.00	29.00	-5.68	0.000*	0.67
		Teacher	16	10.31	165.00				
	3. Measure	Peer	56	43.93	2460.00	32.00	-5.65	0.000*	0.67
		Teacher	16	10.50	168.00				
Online	1. Measure	Peer	56	43.50	2436.00	56.00	-5.32	0.000*	0.63
		Teacher	16	12.00	192.00				
	2. Measure	Peer	56	42.83	2398.50	93.50	-4.81	0.000*	0.57
		Teacher	16	14.34	229.50				
	3. Measure	Peer	56	43.13	2415.50	76.50	-5.04	0.000*	0.59
		Teacher	16	13.28	212.50				
Anonymous	1. Measure	Peer	56	38.85	2175.50	316.50	-1.78	0.075	--
		Teacher	16	28.28	452.50				
	2. Measure	Peer	56	38.35	2147.50	344.50	-1.40	0.161	--
		Teacher	16	30.03	480.50				
	3. Measure	Peer	56	38.99	2183.50	308.50	-1.89	0.058	--
		Teacher	16	27.78	444.50				

P.S. * $p < .05$ Criteria: "Peer=1"; "Teacher=2"; N shows the total scores made, not the number of people (For peer $7 \times 8 = 56$, for teachers $2 \times 8 = 16$).

Table 4 above shows a statistically significant difference between peer and teacher assessments in all three intervals ($p < .05$) for online and face-to-face environments. Considering the effect sizes of these differences, which were found statistically significant, all had a large effect size. However, no statistically significant difference occurs between the peer and teacher assignments in all intervals of the anonymous assessment environment ($p < .05$).

The researchers interviewed all participants to obtain in-depth information about the reason(s) for this situation, focusing on whether students favoured their assessment environments (if yes, why; if no, why). Almost all participants approved anonymous assessments but disapproved of face-to-face and online assessments due to recognising the identity as assessors. Anonymity helps them feel safe to assess their peers, which results in fair and objective assessments and scores. Besides, as there is no peer pressure, they feel comfortable evaluating their peers to give more critical feedback. These findings may explain why peer and teacher assessments were close to each other in the anonymous assessment. However, those who assessed their peers face-to-face and online complained about being known as assessors, which causes peer pressure, unfair and subjective assessment.

Table 5. *Opinions of Participants Related to Assessment Environments*

Codes	f	Representative excerpts
(No) peer pressure	23	I felt compelled to give high scores due to the possible reactions of my peers (FP1) My identity was known, so I couldn't give a low score to avoid peer pressure (OP3) I felt safe to assess my peers; therefore, my score was influenced by my peers (AP8)
(Un)fair assessment (Objective & subjective scoring)	18	It was not a fair process because I could not score my peers in an honest way (FP5) My scoring was subjective because the assesseees were aware of my identity (OP2) Since neither the assessors nor the assesseees knew which scoring was mine, I could make a fair assessment (AP2)
(Un)critical feedback	17	Although I did not like Ali's (pseudonym) teaching, I gave positive and superficial feedback to prevent our friendship from being damaged (FP4) If I were the lecturer or if my identity had not been known, I would have given appropriate feedbacks to the lessons performed by Hakan, Ali and Pelin (pseudonyms) (OP5) Anonymity let me provide substantively critical feedback on the performance of my peers (AP1)
Feel (un)comfortable when evaluating	17	How can I feel comfortable if my friends learn the low scores I gave? (FP6) It is annoying to know that your peers will judge you for not giving high scores (OP7) As no one knew which scoring was mine, I felt quite comfortable during the assessment procedure (AP4)

The assessment environment affects students' assessment. Thus, MFRM was applied to determine rater biases that caused this difference. MFRM was carried out within the scope of a fully crossed design.

As is seen in Table 6 above, the discrimination rate, discrimination index, and discrimination index reliability were high (students were distinguished successfully according to their teaching performance, the raters exhibited different behaviours in performance assessment, there was a difference in assessing the performance of the individual according to the assessment environments, the evaluations in all three time periods were different from each other, the peer and teacher assessment differed in the performance assessment, and the items could correctly distinguish the students' teaching performance). The chi-square values displayed a statistically significant difference ($p < .05$). In other words, the difference between the elements that make up each facet was statistically significant. The chi-square value, standard deviation, discrimination index, and the reliability of the discrimination index predicted for anonymous assessment were lower than face-to-face and online assessments. After analysing the Rasch analysis results for each facet, the rater bias was tried to be determined. Each rater bias is presented in separate headings.

Table 6. Many-Facet Rasch Analysis for Online, Face-to-face and Anonymous Assessments

Assessment Environments	Measurements	Person	Rater	Interval	Rater Type	Item
Online Assessment	Mean of rating observed	3.12	3.14	3.33	2.89	3.12
	Standard deviation of rating observed	0.38	0.35	0.26	0.58	0.23
	Logit minimum value	-1.15	-2.15	-1.70	-1.21	-2.93
	Logit maximum value	4.90	0.96	1.31	1.21	1.79
	Logit mean	2.86	0.00	0.00	0.00	0.00
	Logit standard deviation	2.27	1.05	1.54	1.71	1.35
	RMSE	0.11	0.12	0.07	0.06	0.17
	Discrimination rate	20.69	8.46	22.85	26.50	7.89
	Discrimination index	27.92	11.61	30.80	35.66	10.85
	Chi-square	3 047.70	733.80	1 080.70	703.10	1 157.60
	p-value (for χ^2)	0.00	0.00	0.00	0.00	0.00
Discrimination index reliability	1.00	0.99	1.00	1.00	0.98	
Face-to-Face Assessment	Mean of rating observed	2.96	2.98	2.96	2.70	2.96
	Standard deviation of rating observed	0.36	0.41	0.24	0.68	0.21
	Logit minimum value	-0.80	-2.07	-1.19	-1.14	-2.59
	Logit maximum value	3.77	1.18	1.10	1.14	1.18
	Logit mean	1.53	0.00	0.00	0.00	0.00
	Logit standard deviation	1.81	1.08	1.15	1.61	1.01
	RMSE	0.10	0.11	0.06	0.06	0.15
	Discrimination rate	18.37	9.69	19.05	28.18	6.58
	Discrimination index	24.83	13.25	25.74	37.91	9.10
	Chi-square	2 351.80	954.10	725.90	795.20	794.70
	p-value (for χ^2)	0.00	0.00	0.00	0.00	0.00
Discrimination index reliability	1.00	0.99	1.00	1.00	0.98	
Anonymous Assessment	Mean of rating observed	2.91	2.92	2.91	2.84	2.91
	Standard deviation of rating observed	0.61	0.16	0.37	0.18	0.28
	Logit minimum value	-3.50	-0.57	-2.15	-0.36	-3.58
	Logit maximum value	5.49	0.60	1.88	0.36	2.35
	Logit mean	1.95	0.00	0.00	0.00	0.00
	Logit standard deviation	3.31	0.38	2.03	0.51	1.58
	RMSE	0.11	0.12	0.06	0.06	0.16
	Discrimination rate	31.21	3.10	31.41	8.11	9.69
	Discrimination index	41.94	4.46	42.21	11.15	13.26
	Chi-square	7 116.20	98.90	1 954.80	66.80	1 676.90
	p-value (for χ^2)	0.00	0.00	0.00	0.00	0.00
Discrimination index reliability	1.00	0.91	1.00	0.99	0.99	

RMSE: square root of mean square error

3.2. Rater Severity and Leniency

Group and individual-level statistics were examined during the examination of rater bias. Considering the group-level statistics, discrimination rate, discrimination index, and discrimination index reliability were high. These values indicate that the raters behaved differently while assessing their peers. Then, the logit measure, one of the individual-level statistics was run. Since the logit measure did not have a critical value for severity and leniency, the t-value obtained using logit values was used.

All raters, except one, exhibited severity and leniency biases in the face-to-face assessment; two raters showed severity and leniency biases in the online assessment; three raters displayed severity and leniency biases in anonymous assessment. Teacher assessment had severity bias in all three assessment environments. Specifically, when the t-value of the teacher assessments in the face-to-face and online assessment environments was examined, teachers were observed to be more severe than peers. This might be due to the

leniency bias of peers while assessing each other. Regarding anonymous assessment, the t-value of teacher assessment was much smaller, and there were more peers with neutral behaviours.

Table 7. *T-Values of Raters*

Raters	Face-to-face		Online		Anonymous	
	t-value	p-value	t-value	p-value	t-value	p-value
Peer1	9.83	0.00*	7.38	0.00*	5.00	0.00*
Peer2	7.55	0.00*	4.92	0.00*	3.42	0.01*
Peer3	7.36	0.00*	4.77	0.00*	2.92	0.02*
Peer4	6.09	0.00*	4.46	0.00*	0.58	0.57
Peer5	3.64	0.01*	2.69	0.02*	-0.17	0.87
Peer6	3.18	0.01*	2.83	0.02*	-0.25	0.81
Peer7	-2.27	0.05	2.08	0.07	-0.75	0.47
Peer8	-3.27	0.01*	0.67	0.52	-1.42	0.19
Teacher1	-15.70	0.00*	-15.27	0.00*	-5.00	0.00*
Teacher2	-20.70	0.00*	-19.55	0.00*	-5.18	0.00*

*p < .05; tcritic (0.05;9) = 2.26

One interview question was, “What do you think about the correlation between peer and teacher assessment in terms of scoring?”. Supporting the quantitative results, the qualitative data emphasised that teachers were not severe; on the contrary, students’ leniency bias created this misunderstanding. Students in online and face-to-face environments were inclined to give higher scores due to the reasons above, including recognition of identity, social influence, and relationships based on mutual interests.

Table 8. *The reasons for severity and leniency*

Codes	F	Representative excerpts
Recognition of the identity	16	The situation is not the same for us. For example, I didn’t feel comfortable while assessing because they knew which scoring was mine (OP3) The difference between peer and teacher assessment is due to non-anonymity. The scores given by teachers reflect the quality of the performance. However, our scores symbolise our companionship. (FP2) I usually give high scores to peers as I do not want my value to decrease for them. (FP5)
Social influence	10	I felt inhibited to give low scores during the peer assessment procedure because of the presence or action of peers. (OP1)
Relationship-based on mutual interests	5	... there is a relationship based on mutual interests. If you give high scores, you receive high scores too or vice versa. (OP5) I don’t give low scores to those who gave me high scores (FP5)

3.3. Central Tendency

The first step was to examine category statistics, one of the group-level statistics.

Table 9. *Calculated Category Statistics for Assessment Environments*

Assessment environments	Rating categories	Frequency	%	Cumulative %	Average logit measure	Expected logit measure	Outfit
Face-to-face	1	136	3	3	-4.33	-4.09	0.80
	2	828	20	23	-0.86	-0.99	1.10
	3	2 198	54	77	2.32	2.38	1.10
	4	942	23	100	5.19	5.14	0.90
Online	1	69	2	2	-5.10	-5.10	1.00
	2	614	15	17	-1.08	-1.21	1.20
	3	2 185	53	70	3.09	3.16	0.90
	4	1 236	30	100	6.90	6.84	0.90
Anonymous	1	242	6	6	-5.31	-5.18	0.90
	2	885	22	27	-1.86	-1.79	0.80
	3	1 967	48	75	2.68	-0.51	1.10
	4	1 010	25	100	6.36	5.33	1.20

The most preferred rating category was good (the third one), and the least preferred was poor (the first one). This might be due to the central tendency effect and moderate individual performance. Discrimination rate, discrimination index, and discrimination index reliability concerning individual facet were found high. In

other words, students' teaching performance was determined to be differentiated successfully from each other. In this case, the current situation in the rating categories results from the student's performance. That is, no central tendency effect was determined at the group level for all three assessment environments. The outfit and infit values for the rater facet and the category statistics calculated for each rater were examined to determine whether there was a central tendency effect at the individual level. It was observed that all raters' outfit and infit values were within the acceptable ranges (0.5 and 1.5).

Table 10. *Category Statistics Calculated for Each Rater*

Assessment Environment	Rater	Outfit Statistics				Central Tendency
		Category 1	Category 2	Category 3	Category 4	
Face-to-face	Peer1	-	1.10	1.10	0.90	No
	Peer2	-	1.10	0.90	0.90	No
	Peer3	-	0.90	1.20	1.00	No
	Peer4	-	1.00	1.00	0.90	No
	Peer5	1.70	1.70	1.30	1.10	Yes
	Peer6	1.00	0.90	0.70	0.90	No
	Peer7	-	0.90	1.10	0.90	No
	Peer8	1.10	0.90	1.00	1.00	No
	Teacher1	0.80	1.20	1.20	0.90	No
	Teacher2	0.70	0.90	1.10	1.00	No
Online	Peer1	-	1.40	0.70	0.70	No
	Peer2	-	1.70	0.70	0.70	Yes
	Peer3	-	1.10	0.60	0.80	No
	Peer4	-	1.60	0.90	0.60	Yes
	Peer5	-	2.10	1.10	1.10	Yes
	Peer6	-	1.50	1.40	1.40	No
	Peer7	1.20	1.30	0.70	0.90	No
	Peer8	-	0.60	0.80	0.80	No
	Teacher1	1.30	1.30	1.20	1.10	No
	Teacher2	0.90	1.00	0.90	0.70	No
Anonymous	Peer1	0.60	1.50	0.90	1.00	No
	Peer2	1.00	1.00	1.70	1.20	Yes
	Peer3	1.00	0.60	0.90	1.20	No
	Peer4	0.80	0.70	1.20	1.20	No
	Peer5	0.80	0.60	1.00	1.10	No
	Peer6	1.50	0.80	1.20	1.40	No
	Peer7	1.00	0.90	0.90	1.20	No
	Peer8	0.90	0.70	1.00	1.50	No
	Teacher1	0.60	0.80	1.40	1.50	No
	Teacher2	0.70	0.70	0.80	1.00	No

One of the raters in the face-to-face assessment (Peer5), three in the online assessment (Peer2, Peer4, and Peer5), and one in the anonymous assessment (Peer2) exhibited a central tendency effect when assessing students' teaching performance. Although all anonymous raters preferred the first category, it was ignored by the majority in the other assessment environments. Since the identity of the raters in face-to-face and online assessment environments were known, they could not give a "poor" rating. Interviews confirmed the quantitative data, emphasising that recognising assessors' identity and personality clash might cause central tendency.

Table 11. *The Reasons for Central Tendency*

Codes	f	Representative excerpts
Recognition of the identity	16	Even I did not like my peers' performance, I usually tick "fair" instead of "poor" because they knew it was my score. (OP5)
		I felt inhibited to give low scores during the peer assessment procedure because of the presence or action of peers. (OP1)
Personality clash	10	I lowered a score when I didn't like a peer. (FP8)
		If a peer had disturbing behaviour, I lowered his/her score. (FP6)

3.4. Halo Effect

During the performance assessment process, the measurement report related to the item facet was investigated to determine the halo effect at the group level. Discrimination rate, discrimination index, and discrimination index reliability were found high. These high values indicate that the items had different performance levels and successfully differentiated the individual's performance from each other so that the halo effect did not interfere with the ratings. On the other hand, no halo effect was observed at the group level. Thus, the difference between the calculated logit values for rating criteria was examined, and this difference was found to be significantly higher than one in three assessment environments. In this context, raters whose fit statistics were statistically higher than one are stated to exhibit the halo effect (MyFord & Wolfe, 2004).

Table 12. *The Infit and Outfit Fit Values for the Raters*

Rater	Face-to-face				Online				Anonymous			
	Infit MnSq	Zstd	Outfit MnSq	Zstd	Infit MnSq	Zst d	Outfit MnSq	Zstd	Infit MnSq	Zstd	Outfit MnSq	Zstd
Peer1	1.06	0.8	1.07		1.00		0.95	-0.3	1.06	0.8	1.10	0.7
Peer2	0.98	-0.2	0.99		0.85		0.77	-2.1	1.01	0.2	0.98	-0.1
Peer3	0.97	-0.4	0.91		1.21		1.26	2.1	1.11	1.4	1.30	2.1
Peer4	0.93	-1.1	0.88		0.81		0.73	-2.6	1.06	0.7	1.04	0.3
Peer5	0.99	0.0	0.99		0.81		0.73	-2.6	1.14	1.8	1.18	1.6
Peer6	0.96	-0.5	0.95		0.86		0.76	-2.4	0.99	-0.1	0.96	-0.3
Peer7	1.01	0.1	0.97		1.28		1.31	2.7	0.93	-0.9	0.90	-0.8

The values that raters received did not differ from one in all three assessment environments. For the final decision whether there is a halo effect at the individual level, the item difficulties were equalised, and raters whose infit and outfit values were equal to one were assumed to show the halo effect. After balancing the item difficulties for each assessment environment and examining the raters' infit and outfit values, only a peer (peer7) exhibited the halo effect in the online assessment.

Students answered the following questions during the interviews: "Does the overall impression of a peer impact your assessment of that peer's performance? Is there a relationship between one's attractiveness and the quality of his/her teaching?" The participants' statements justified the quantitative findings, indicating that students did not have a halo effect.

During the peer assessment, as you can see on Edmodo, I only focused on the quality of peers' teaching. OP4

We were asked to assess our peers' teaching performance, not personality. Thus, whether the assessee was an attractive peer or a close friend did not interfere in my scoring. AP6.

4. Conclusion and Discussion

This paper aimed to examine the rater biases involved in the measurements when their peers and teachers assessed the teaching performance of students and to determine the reasons for these biases. The ARP developed by the researchers to assess students' teaching performance was found to provide valid and reliable measurements. Thus, it is recommended that further studies use the ARP when examining the teaching performance of students.

A statistically significant difference was found between the teacher and peer raters in the online and face-to-face assessment environments. In contrast, no statistically significant difference was found between anonymous teachers and peer raters. Besides, the obtained significant difference was noted to have a large effect size. Considering the intervals, the difference between the teacher and peer raters was statistically significant in the online and face-to-face assessment environments, but it was insignificant in anonymous assessment. The qualitative data were performed to examine why teachers and peers had similar scoring in anonymous assessment and anonymity was found to be an important factor for this situation, confirming the previous research (Cheng & Tsai, 2012; Chester & Gwynne; Pope, 2005; Yu & Sung, 2016; Vickerman, 2009). Thus, anonymity provides a safe environment for peer raters, which results in fairer assessment as there is no over-scoring (Freeman & McKenzie, 2000; Panadero et al., 2013).

The severity and leniency biases were observed during teaching performance assessment at individual and group levels, confirming various studies (Knoch et al., 2018). This shows that the severity and leniency biases

are important behaviours in rater inconsistency (Kane et al., 1995). Considering the severity and leniency biases at the individual level, the anonymous assessment was found to have the least bias. In contrast, the online and face-to-face assessments displayed similar biases. According to qualitative data, this is due to the recognition of their identity as assessors. In other words, when the assessors' identity is known, friendship comes to the fore rather than the actual performance of the individuals. As the assessor was not known in the anonymous assessment, few peer raters showed the severity and leniency biases. It is stated in the literature that peers are stressed during the scoring process when their identities are known (Pope, 2005; Yu & Sung, 2016).

Central tendency is another error involved in measurements originating from the raters when assessing the performance of the individuals. No central tendency effects were found at the group levels in this research. The literature supports this result. The study conducted by Esfandiari (2015) emphasises that some of the raters showed the central tendency effects during the assessment of the academic writing skills at the individual level but not at the group level. Also, the central tendency effects appeared less in performance assessment compared to the severity and leniency biases. This result shows that the most common behaviours in performance assessment are the severity and leniency biases (Cronbach, 1990). Considering the central tendency effects at the individual level, one rater preferred a certain category of the measurement tool more than other raters in the face-to-face and anonymous assessments, whereas three raters did in online assessment. When qualitative interviews were analysed for why all three raters preferred the second category and never chose the first category in online evaluation, they were observed to avoid choosing the first category and favour the second category as the identity of peer assessors was known. This confirms Yu and Sung (2016), who advocates that students' feedback will vary depending on whether their identities as assessors are revealed.

Another rater bias frequently is the halo effect. Farrokhi and Esfandiari (2011) examined the interference of the halo effect in performance during the process of peer, self, and teacher assessment. They found that the halo effect occurred in all three evaluation types. However, the current study found no halo effect at the group and individual levels (except one rater) regarding all three evaluation environments. This may be because students were knowledgeable about the halo effect and its consequences. As Myford and Wolfe (2004) suggested, raters should be informed about the halo effect and how it affects the scores.

When peer raters participate in the scoring process (assessment), the validity and reliability of the scoring become a major concern (May, 2008). Therefore, evidence must be gathered for the validity and reliability of the assessments. In this context, in the current study, evidence was collected for the validity of the measurements by examining the rater biases involved in the measurements during the peer assessment process. The current study provided the following results:

- The recognition of the identity was effective on scoring when peers assessed the teaching performance of students. It was also found that there was a difference between the scores made over time.
- The severity and leniency biases were observed in the peer assessment process at the individual and group levels. In addition, peer raters exhibited more leniency biases instead of severity biases. The anonymous assessment was found to have the least bias, while the online and face-to-face assessments exhibited similar biases.
- There were no central tendency and halo effects at the group level, but they were found to interfere with the measurement at the individual level.

Based on the findings of the current study, the following recommendations were provided:

- During the performance assessment process, rater training can be arranged to reduce the rater biases.
- Anonymity in the peer assessment process can contribute to the validity and reliability of the measurements.
- Considering the similarity of scores in face-to-face and online assessment environments, either of them can be used in the peer assessment process.

As can be seen from the results of this research, rater biases are inevitable in the performance evaluation process. In this context, further research can focus on the effectiveness of rater training and investigate whether rater training will effectively reduce these biases. While explaining the rater bias patterns of EFL students, this study ignored the gender variable because the majority were females. Further studies may investigate whether females and males differ in terms of severity and leniency.

5. Limitations

Research has some limitations. First, the findings may not be generalised as the study was conducted with students in the English language teaching department. Another limitation is that only four rater biases most involved in the measurements were considered in this study, although peer raters have many rater biases when evaluating individual performance. Another limitation is that the raters in this research are novice/inexperienced. The last limitation is that students' teaching performance was examined, but their performance in other lessons was not considered.

6. References

- Abu Kassim, N.L. (2007). Exploring rater judging behaviour using the many-facet Rasch model. *Paper Presented in 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/3212/>
- Anastasi, A. (1976). *Psychological testing* (4th ed.). Macmillan.
- Akpinar, M. (2019). The effect of peer assessment on pre-service teachers' teaching Practices. *Education & Science*, 44(200), 269-290. <https://doi.org/10.15390/EB.2019.8077>
- 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*. University of Oxford for Educational Assessment. Retrieved from <https://dera.ioe.ac.uk/17683/1/2013-01-21-marker-effects-and-examination-reliability.pdf>
- Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. *The companion to language assessment*, 3, 1301-1322. <https://doi.org/10.1002/9781118411360.wbcla070>
- Bennett, J. (1998). *Human resources management*. Prentice Hall.
- Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge. <https://doi.org/10.4324/9781315814698>
- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89-110. <https://doi.org/10.1191/0265532203lt245oa>
- Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, 41(3), 400-413. <https://doi.org/10.1080/02602938.2015.1018133>
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL quarterly*, 32(4), 653-675. <https://doi.org/10.2307/3587999>
- Cetin, B., & Ilhan, M. (2017). An analysis of rater severity and leniency in open-ended mathematic questions rated through standard rubrics and rubrics based on the SOLO taxonomy. *Education and Science*, 42(189), 217-247. <https://doi.org/10.15390/EB.2017.5082>
- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265-289. <https://doi.org/10.2307/1165285>
- Cheng, K. H., & Tsai, C. C. (2012). Students' interpersonal perspectives on, conceptions of and approaches to learning in online peer assessment. *Australasian Journal of Educational Technology*, 28(4), 599-618. <https://doi.org/10.14742/ajet.830>
- Chester, A., & Gwynne, G. (2006). Online teaching: encouraging collaboration through anonymity. *Journal of Computer-Mediated Communication*, 4(2), JCMC424. <https://doi.org/10.1111/j.1083-6101.1998.tb00096.x>

- Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328-338. <https://doi.org/10.1016/j.learninstruc.2009.08.006>
- Cronbach, L.I. (1990). *Essentials of psychological testing*. Harper and Row Publishers.
- Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112. <https://doi.org/10.1111/j.1745-3984.1994.tb00436.x>
- Esfandiari, R. (2015). Rater errors among peer-assessors: applying the many-facet Rasch measurement model. *Iranian Journal of Applied Linguistics, 18*(2), 77-107. <https://doi.org/10.18869/acadpub.ijal.18.2.77>
- Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory & Practice in Language Studies, 1*(11), 1531-1540. <https://doi.org/10.4304/tpls.1.11.1531-1540>
- Farrokhi, F., Esfandiari, R., & Vaez Dalili, M. (2011). *Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment*. *World Applied Sciences Journal, 15*(11), 76-83.
- Freeman, M., & McKenzie, J. (2000). Self and Peer Assessment of Student Teamwork: Designing, implementing and evaluating SPARK, a confidential, web based system. *Flexible learning for a flexible society*. Retrieved from <https://ascilite.org/archived-journals/aset/confs/aset-herdsa2000/procs/freeman.html>
- Goodrich, H. (1997). Understanding Rubrics: The dictionary may define" rubric," but these models provide more clarity. *Educational Leadership, 54*(4), 14-17.
- Güneş, P., & Kiliç, D. (2016). Dereceli puanlama anahtarı ile öz, akran ve öğretmen değerlendirmesi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 1*(39), 58-69. <https://doi.org/10.21764/efd.93792>
- Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics, 33*(2), 87-102.
- Haladyna, T. M. (1997). *Writing test items in order to evaluate higher order thinking*. Allyn & Bacon.
- Hansen, K. (2003). *Writing in the social sciences: A rhetoric with readings*. Pearson Custom.
- Hosack, I. (2004). The effects of anonymous feedback on Japanese university students' attitudes towards peer review. In R. Hogaku (Ed.), *Language and its universe* (pp. 297-322). Ritsumeikan Hogaku.
- Ilhan, M. (2016). A Comparison of the Ability Estimations of Classical Test Theory and the Many Facet Rasch Model in Measurements with Open-ended Questions. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 346-368.
- Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field methods, 18*(1), 3-20. <https://doi.org/10.1177/1525822X05282260>
- Kane, J., Bernardin, H., Villanueva, J., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal, 38*, 1036-1051. <https://doi.org/10.2307/256619>
- Khaatri, N., Kane, M.B., & Reeve, A.L. (1995). How performance assessments affect teaching and learning. *Educational Leadership, 53*(3), 80-83.
- 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>
- Kingsbury, F. A. (1922). Analyzing ratings and training raters. *Journal of Personnel Research, 1*, 377-383.
- Knoch, U., Fairbairn, J., Myford, C., & Huisman, A. (2018). Evaluating the relative effectiveness of online and face-to-face training for new writing raters. *Papers in Language Testing and Assessment, 7*(1), 61-86.
- Kutlu, Ö., Doğan, C.D., & Karaya, İ. (2014). Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme [Determining student success: Determination based on performance and portfolio]. Pegem Akademi Yayıncılık.

- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, 28(4), 563-575. <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
- Lawshe, C. H. (1985). Inferences from personnel tests and their validity. *Journal of Applied Psychology*, 70(1), 237-238. <https://doi.org/10.1037/0021-9010.70.1.237>
- Li, L. (2017). The role of anonymity in peer assessment. *Assessment & Evaluation in Higher Education*, 42(4), 645-656. <https://doi.org/10.1080/02602938.2016.1174766>
- Li, L., & Gao, F. (2016). The effect of peer assessment on project performance of students at different learning levels. *Assessment & Evaluation in Higher Education*, 41(6), 885-900. <https://doi.org/10.1080/02602938.2015.1048185>
- Li, L., Liu, X., & Zhou, Y. (2012). Give and take: A re-analysis of assessor and assessee's roles in technology-facilitated peer assessment. *British Journal of Educational Technology*, 43(3), 376-384. <https://doi.org/10.1111/j.1467-8535.2011.01180.x>
- Linacre, J. M. (2017). *A user's guide to FACETS: Rasch-model computer programs*. MESA Press.
- Mackay, A., & Gass, S. (2005). *Second Language Research: Methodology and Design*. Lawrence Erlbaum Associates.
- May, G. L. (2008). The effect of rater training on reducing social style bias in peer evaluation. *Business Communication Quarterly*, 71(3), 297-313. <https://doi.org/10.1177/1080569908321431>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-256. <https://doi.org/10.1177/026553229601300302>
- Miyazoe, T., & Anderson, T. (2011). Anonymity in blended learning: who would you like to be?. *Journal of Educational Technology & Society*, 14(2), 175-187.
- Moore, B.B. (2009). Consideration of rater effects and rater design via signal detection theory. (Unpublished Doctoral dissertation). Retrieved from <http://www.proquest.com/>
- Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research, and Evaluation*, 7(1), 3.
- Myford, C. M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education*, 15(2), 187-215. https://doi.org/10.1207/S15324818AME1502_04
- 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.
- Newby, D., Allan, R., Fenner, A. B., Jones, B., Komorowska, H., & Soghikyan, K. (2007). *European Portfolio for Student Teachers of Languages: A reflection tool for language teacher education*. Council of Europe.
- Özdemir, O., & Erdem, D. (2017). Sunum becerilerinin akran değerlendirmesine arkadaşlığın etkisi. *Turkish Journal of Educational Studies*, 4(1), 21-43.
- Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: a review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Human factors and social conditions of assessment* (pp. 1-39). Routledge.
- Panadero, E., Romero, M., & Strijbos, J-W (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, 39(4), 195-203. <https://doi.org/10.1016/j.stueduc.2013.10.005>
- Papinczak, T., Young, L., Groves, M., & Haynes, M. (2007). An analysis of peer, self, and tutor assessment in problem-based learning tutorials. *Medical teacher*, 29(5), e122-e132. <https://doi.org/10.1080/01421590701294323>
- Pope, N. K. L. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education*, 30(1), 51-63. <https://doi.org/10.1080/0260293042003243896>

- Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, 94(1), 31-37. <https://doi.org/10.5951/MT.94.1.0031>
- Rotsaert, T., Panadero, E., & Schellens, T. (2018). Anonymity as an instructional scaffold in peer assessment: its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *European Journal of Psychology of Education*, 33(1), 75-99. <https://doi.org/10.1007/s10212-017-0339-8>
- Royal, K. D., & Hecker, K. G. (2016). Rater errors in clinical performance assessments. *Journal of veterinary medical education*, 43(1), 5-8. <https://doi.org/10.3138/jvme.0715-112R>
- Schools, C. C. P., & Chesterfield, V. (2015). Performance evaluation handbook for teachers. Regina, SK. <https://www.nctq.org/dmsView/70-07>
- Schoonenboom, J., & Johnson, R. B. (2017). How to construct a mixed methods research design. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 69(2), 107-131. <https://doi.org/10.1007/s11577-017-0454-1>
- Sudweeks, R. R., Reeve, S. & Bradshaw, W.S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9, 239-261. <https://doi.org/10.1016/j.asw.2004.11.001>
- Sung, Y. T., Chang, K. E., Chang, T. H., & Yu, W. C. (2010). How many heads are better than one? The reliability and validity of teenagers' self-and peer assessments. *Journal of Adolescence*, 33(1), 135-145. <https://doi.org/10.1016/j.adolescence.2009.04.004>
- Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing*, 34(1), 3–22. <https://doi.org/10.1177/0265532215594830>
- Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., & Schellens, T. (2015). What if pupils can assess their peers anonymously? A quasi-experimental study. *Computers & Education*, 81, 123–132. <https://doi.org/10.1016/j.compedu.2014.10.001>
- Vickerman, P. (2009). Student perspectives on formative peer assessment: an attempt to deepen learning?. *Assessment & Evaluation in Higher Education*, 34(2), 221-230. <https://doi.org/10.1080/02602930801955986>
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287. <https://doi.org/10.1177/026553229801500205>
- Welsh, E. (2002, May). Dealing with data: Using NVivo in the qualitative data analysis process. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 3(2). Retrieve from <http://www.qualitative-research.net/index.php/fqs/article/view/865/1881>
- Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, 45(3), 197-210. <https://doi.org/10.1177/0748175612440286>
- Yu, F. Y., & Liu, Y. H. (2009). Creating a psychologically safe online space for a student-generated questions learning activity via different identity revelation modes. *British Journal of Educational Technology*, 40(6), 1109-1123. <https://doi.org/10.1111/j.1467-8535.2008.00905.x>
- Yu, F. Y., & Sung, S. (2016). A mixed methods approach to the assessor's targeting behavior during online peer assessment: effects of anonymity and underlying reasons. *Interactive learning environments*, 24(7), 1674-1691. <https://doi.org/10.1080/10494820.2015.1041405>