Id: 1300980

Participatory Educational Research (PER)
Vol.10(5), pp. 98-118, September 2023
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.23.77.10.5

The Role of Time on Performance Assessment (Self, Peer, And Teacher) in Higher Education: Rater Drift

Hikmet ŞEVGİN*
Faculty of Education, Department of Educational Sciences, Department of Measurement and Evaluation in Education, Van Yuzuncu Yil University, Van, Türkiye
ORCID: 0000-0002-9727-5865

Mehmet ŞATA
Faculty of Education, Department of Educational Sciences, Department of Measurement and Evaluation in Education, Van Yuzuncu Yil University, Van, Türkiye
ORCID: 0000-0003-2683-4997

Article history
Received:
23.05.2023

Received in revised form:
01.07.2023

Accepted:
08.08.2023

Key words:
Many-facet rasch; presentation skills; rater drift; reliability; validity.

This study aimed to investigate the change in teacher candidates' oral presentation skills over time through self, peer, and teacher assessments using the rater drift method. A longitudinal descriptive research model was used as a quantitative research approach to achieve this aim. The study group consisted of 47 teacher candidates receiving formation education at a state university in the Eastern Anatolia Region and an instructor teaching the course. An analytical rubric was used as a data collection tool to evaluate the candidates' oral presentation skills. The data collection process lasted six weeks in total. Since the performance evaluation process aimed to examine the change over time, the many-facet Rasch model was used. When the findings of the study were examined, it was determined that the rater behavior of teacher candidates had statistically significant differences at the group level over time. It was found that 26 out of 48 peer raters had rater drift in their evaluations. It was also found that the majority of rater drift over time was positive, meaning that evaluators became more generous over time. Another result obtained in the study was that teacher assessment did not show rater drift over time, with similar ratings for six weeks. The study's findings were discussed with previous studies in the literature, and recommendations were made to researchers.

Introduction

Assessment is a systematic process, a fundamental component of educational and instructional services, as it helps identify students' strengths and areas for improvement. The process begins with identifying learning outcomes and ends with determining whether or not those outcomes have been achieved (Linn, 2008; Petra & Ab Aziz, 2020). In addition to evaluating students' academic achievements, contemporary assessment practices involve students taking an active role in their own learning and assessment processes, including

* Correspondency: hikmetsevgin@gmail.com

assessing their various skills and qualities (Orlova, 2019). Studies have shown that involving students in the assessment process can increase learning outcomes and improve the quality of assessments (Boud & Falchikov, 2006).

It can be argued that traditional student assessment approaches are primarily focused on measuring academic achievement and are often evaluative in nature, with the evaluator typically being the teacher and students' skills and qualities beyond their academic achievements being overlooked (Modarresi et al., 2021). Therefore, it is important to assess students' academic achievements, skills, qualities, personal development, and social responsibilities (Dishon & Gilead, 2020; Wayda & Lund, 2005). Traditional assessment approaches are based on evaluating a series of test responses to measure academic achievement. However, all outputs obtained from students, including their responses to test items, should be considered and evaluated using complementary assessment approaches that consider a broader meaning for the student (Board of Education, 2005). Therefore, the use of complementary assessment approaches alongside traditional assessment approaches is intended to improve the quality of classroom measurement and evaluation practices (Dikli, 2003; Kutlu et al., 2010; Oren et al., 2014; Sad & Goktas, 2013; Shepard, 2000). Indeed, the versatility of classroom measurement and evaluation practices is necessary for today's era of new developments and students equipped with advanced knowledge and skills (Szőkol et al., 2022). It is also a fact that a student's potential can only be realized through quality education and multiple performance evaluators. Therefore, in this study, teacher candidates' presentation skills (individually or in groups) were evaluated using self, peer, and teacher assessments together to provide a comprehensive analysis of performance assessment methods over a period of six weeks. In addition, changes in rater behavior over time were also examined.

Performance Assessment

Performance assessment is not a new concept in education (Kutlu et al., 2010). In Turkiye, it gained great importance with transitioning to a constructivist approach that was considered student-centered education in 2006 (Cepni, 2010). Performance assessment helps students apply their academic knowledge and skills to real-life problems (Guler, 2012). Evaluating student performance requires a meticulous assessment that systematically collects evidence of learning (Dunn & Mulvenon, 2009). Furthermore, it can motivate students to achieve success and positive recognition (Szőköl et al., 2022). Performance assessment also develops students' self-monitoring and self-assessment skills (Szőkol et al., 2022). This enables students to demonstrate many skills, such as applying what they have learned, creative thinking, problem-solving, collaboration, and presentation. In other words, it allows for tracking students' academic achievements and their personal and social development. Black and Wiliam (1998) noted that teachers use performance assessments to identify their student's weak points and provide additional help in the areas where they fall short. Similarly, performance assessment helps students understand their learning processes, identify their strengths and weaknesses, learn what they need to do to achieve their goals, and be more effective in the learning process (Maier et al., 2020). In addition to monitoring student performance, performance assessment is also essential in measuring the quality of education teachers provide (Gomleksiz et al., 2011). Performance tasks, observation, drama, student product portfolios, self, peer, and teacher assessment methods are used to assess students' skills and qualities as student performance indicators.

Self, Peer and Teacher Assessment

It is not an easy transition for educators who use traditional assessment approaches to switch to and effectively use complementary assessment approaches. However, complementary approaches are necessary for an effective measurement and evaluation process. According to the research results of Alaz and Yarar (2009), teachers often prefer traditional assessment approaches. They also prefer complementary assessment approaches such as performance tasks, observation, and drama methods. Especially after the Turkish education system shifted to a constructivist approach, traditional assessment approaches started to give way to complementary assessment approaches (Kostereliogl u & Celen, 2016). Similarly, according to Yurdabakan (2012), a structural change in education programs in Turkey has increased interest in complementary assessment approaches. However, according to Duban and Kucukyilmaz (2008), there are still some ongoing problems in using complementary assessment approaches in primary schools.

It can be stated that teacher, peer, and self-assessment have emerged as complementary assessment approaches. According to Duban and Kucukyilmaz (2008), the most frequently used complementary assessment approaches are student product portfolios and performance tasks. The project works and rubrics are occasionally used in group activities. However, concept maps, self-assessment, and peer assessment are less commonly used. Kostereliogl u and Celen (2016) note that teachers are within the self-assessment method based on the constructivist approach but are unwilling to prefer some methods and evaluations.

It can be observed that complementary assessment approaches have a comprehensive structure, allowing both process and product to be evaluated. In other words, it can be stated that complementary assessment approaches have some benefits. According to Erman-Aslanoglu (2022), self and peer assessment make students' evaluations more formal and systematic than the current measurement and evaluation approach. Kilic and Gunes (2016) suggest that despite some limitations of self and peer assessment practices, they have positive features such as motivating students, making them aware of their learning, and improving the quality of their learning process and products. According to Arik and Kutlu's (2013) research, self, and peer assessment are important competence areas for teachers. Tunkler's (2019) research shows that teacher candidates have positive views regarding peer assessment. Ozpinar's (2021) research demonstrates that teacher candidates' self-efficacy regarding the teaching process increases with self, peer, group, and instructor evaluations. According to Kostereliogl u and Celen's (2016) research, teacher candidates have a positive attitude towards using self-assessment methods.

Traditional measurement and evaluation approaches are not practical for group work. Similarly, according to Erman-Aslanoglu (2017), the most important problem in group work is how to evaluate it. Research indicates that both process and product should be evaluated. In process evaluation, self, peer, and teacher assessments stand out. According to Kostereliogl u and Celen (2016), in self-assessment, the student compares their current work with their previous work and asks questions about how much progress they have made, but self-assessment alone is insufficient. Peer and teacher evaluations are also necessary. According to Yildiz (2018), self-assessment can be considered an alternative assessment tool that creates awareness of the student's learning level. According to Uzun and Yurdabakan (2011), the fear of giving oneself a high score is one of the common views in self-assessment. In addition, according to Kilic and Gunes (2016), graded scoring keys are frequently used in performance assessments, and both peer and self-assessment play an important role in educational research.

The presence of problematic situations, such as students' lack of objectivity in the self and peer assessment process or their tendency to overestimate or underestimate themselves and their peers, that are influenced by personal relationships, affects the validity and reliability of measurements (Alici, 2010; Gelbal & Kelecioglu, 2007; Gocer et al., 2017). Therefore, teachers must provide proper guidance and training to students to enable them to evaluate themselves and their friends objectively. The use of scoring criteria (Rubric) and ensuring that the criteria used are transparent and fair will ensure that evaluations are accurate and fair, thus providing the necessary evidence to collect for the reliability and validity of the scores (Donnon et al., 2013; Hafner & Hafner, 2003). Furthermore, the involvement of multiple raters, especially in teacher evaluations, will provide evidence for increasing reliability and validity by ensuring assessor reliability (Karakaya, 2015).

Rater Drift

The process of conducting accurate scoring can be described as a complex process. Errors in scoring can arise during the process, leading to the rater effect. Biases in evaluation results from raters threaten the validity of measurements since they are sources of irrelevant-construct variance being measured (Mesick, 1995). Valid measurements of a student's performance through scoring by different raters can only be achieved if the scoring is reliable (Hafner & Hafner, 2003; Farrokhi et al., 2011; Nalbantoglu Yilmaz, 2017). However, various factors related to raters can mix with measurement results when evaluating a student's performance (Erman-Aslanoglu & Sata, 2023). These factors related to raters that affect a student's performance are referred to as the rater effect (Farrokhi et al., 2011). The decisions of raters, or the rater effect, directly affect the fairness and reliability of students' scores. In the literature, there are various sources of error related to raters. The most commonly encountered ones in the literature are scorer severity and leniency, the halo effect, central tendency bias, and rater drift (Dogan & Uluman, 2017; Hoyt, 2000; Szőkol et al., 2022; Erman-Aslanoglu & Sata, 2023; Wolfe et al., 2007).

The term "rater characteristic" encompasses rater severity and other rater effects (McNamara & Adams, 1991). Generally, rater drift is defined as changes in rater behavior over time. Additionally, the literature indicates that rater drift is inevitable (Park, 2011). Rater drift is a problem that can arise in evaluating student performance, especially in the assessment of assignments or exams used to evaluate student performance in education. Rater drift refers to the phenomenon in which evaluators' judgment standards used in a measurement task change over time or in different contexts, resulting in inconsistent or unreliable ratings (Case, 1997; Harik et al., 2009). As a result, rater drift can lead to inconsistency or bias in ratings, jeopardizing the reliability and validity of the measurement.

Although defined in various ways, rater drift is a concept that has been studied by many researchers (Lamprianou, 2006). Rater drift is a problem, especially when there are a large number of groups to be rated. Systematic effects such as drift in criteria, fatigue, and quality range can contribute to rater drift (McLaughlin et al., 2009; Quellmalz, 1980). In addition, the literature on the use of statistical methods to correct rater or task effects is limited (Raymond et al., 2011). Furthermore, according to Park (2011), although many studies define rater drift as a problem, its impact on the model has not been examined based on a model, and the literature on rater drift tends to focus more on rater severity.

The term "rater drift" refers to a condition stemming from the rater himself/herself or different factors. This condition is a source of concern in studies related to performance tasks,

including cognitive tests (Kooken et al., 2017). It poses a significant threat to performance evaluation (Wesolowski et al., 2017). Researchers can take various measures to minimize these threats. These include monitoring ratings to determine which rater deviates when and where during the scoring process, retraining individual raters, or requiring a re-evaluation of responses from students suspected of being affected by rater drift (Haladyna & Rodriguez, 2013; Wolfe et al., 1999). Pre-scored responses can also be randomly distributed to raters for re-scoring to help identify rater drift (Haladyna & Rodriguez, 2013). However, considering that raters are human (Guilford, 1936), it is impossible to guarantee that scoring judgments will not change over time (Congdon & McQueen, 2000; Harik et al., 2009; Mesick, 1995). Therefore, research in this area is essential, and despite the application of standard measures such as rater retraining and frequent feedback, vigilance is required regarding rater drift.

In this study, the six-week performance of teacher candidates was evaluated and rated by themselves, peers, and teachers, and rater errors related to these ratings were examined. During the implementation phase of the study, a one-week rater training was provided, and a graded scoring rubric was used to obtain more reliable measurements. The Data analysis stage used the many-Facet Rasch Measurement (MFRM) approach. MFRM is recommended for determining the reliability and validity of self, peer, and teacher assessment ratings (Erman-Aslanoglu & Sata, 2023; Uto, 2022). It can also be said that MFRM proposes various extensions to investigate the phenomena of time-specific rater severity, also known as rater drift (Uto, 2022). The main advantages of MFRM include (a) estimation of parameters independent of both test takers and test items, (b) consideration of all facets involved in calibration, (c) calibrated surfaces sharing the same metric, and (d) additivity of estimation values (Zhu & Cole, 1996). Additionally, MFRM is known to be very useful in analyzing the interaction between raters and determining whether any biases are apparent when they score/evaluate different versions of a test (Wigglesworth, 1994). All these aspects make MFRM a suitable option for performance evaluations affected by rater-related behaviors (Mulqueen et al., 2000).

When the literature is examined, it is seen that numerous studies tend to depict rater effects as static properties of raters (i.e., as if a rater effect affects the performance of each student in precisely the same way) in the self-, peer-, and teacher assessments of student performance (Engelhard & Myford, 2003; Harik et al., 2009; Hoskens & Wilson, 2001; Lamprianou, 2006; Leckie & Baird, 2011; McLaughlin et al., 2009; Sata & Karakaya, 2022; Uto, 2022). However, the behavior of an individual rater can change over time, and differential rater functioning over time (drift) can express this change in scorer performance over time. In other words, drift can show that the severity levels of raters systematically change over time (Myford & Wolfe, 2009). It can be said that there are limited studies on whether the effects of the scores given by raters in longitudinal studies are constant over time or whether they change, and their findings vary. For example, Borkan (2017) investigated the degree of rater severity drift in peer assessment among education faculty students. In her study, she reported that 29 students participated in peer assessment at a four-day interval, and peer raters scored their peers quite generously. She also reported that raters' severity/generosity levels differed and became even more severe over time, with the highest severity being on the fourth day. Likewise, Congdon and McQueen (2000) used MRFM to determine the predictions of 16 raters who scored the writing performance of 8,285 primary school students for seven days by two trained raters. In their study, in the analyses of rater frequency estimates (rater stringency estimates), they observed that relative rater severity (the severity of a scorer compared to other scorers on the same day) changed from day to day without a seemingly predictable model and that ten raters differed significantly from their initial predictions towards the end,

with nine raters becoming even more severe and one rater becoming more generous.

Some researchers have stated that the result of rater drift is meaningless. For instance, Leckie and Baird (2011) examined the presence of rater severity effects and central tendency effects among raters who scored the English writing performance of 14-year-old students in their study. Although less experienced raters appeared more stringent than more experienced raters, this result was not statistically significant. However, they reported a central tendency in the raters' scoring and that rater severity was significantly unstable over time. In their study with 7th-grade middle school students, Erman-Aslanoglu and Sata (2023) used MRFM analysis to investigate the temporal shift of rater agreement in peer evaluation for eight oral presentations based on group work in Science and Technology class. They attempted to determine whether the raters shifted their agreement on a group or individual basis by calculating two separate indices (interaction term and standardized differences). As a result, they found that the two methods used to determine rater agreement shift yielded similar results. While there was no significant rater shift at the group level, some raters tended to be more strict or lenient over time at the individual level. There was no specific pattern of shifts.

Importance and Purpose of the Research

Given the effects of rater effects on both the validity and reliability of measurements, it is important to explore these effects. Considering that the effects of rater effects in the performance appraisal process are examined in this study, it is seen that the research is important and contributes to the literature. Another feature that makes this study important is that both peer, self- and teacher assessments were examined at the same time. Considering teacher, peer and self-assessment together provides evidence for the validity and reliability of the measurements and enables the determination of the source of variance that affects the measurements the most. In addition, the fact that this study is a longitudinal study is also considered very important, so that it is possible to determine which rater has what kind of effect on the measurements over time. In the context of determining the role of time in performance assessment, it can be said that the research contributes to the field and is original.

The study aims to examine the emerging rater behaviors in the context of self, peer, and teacher assessments during the evaluation of teacher candidates' presentation performances over time. The problem statement of the research is formulated as follows: Are significant drifts observed in the raters' ratings during the time-based performance assessment process?. Based on the research problem, the following questions have been addressed:

(1) Are there any noticeable drifts in the raters' ratings during the time-based assessment?
(2) How do these drifts impact the overall reliability and validity of the performance assessment?

Method

In the method section, the research model, study group, data collection process, data collection tools and data analysis process are given in detail and evidence for the reliability and validity of the research is presented.

Research Model

This study aimed to examine the changes over time in self, peer, and teacher assessments during the assessment of teacher candidates' presentation skills. To achieve this goal, the study was conducted using a longitudinal descriptive research model, a quantitative research approach (Sata, 2020a). In the study, teacher candidates' group or individual presentations were scored by themselves, their peers, and the course instructor for 6 weeks, and a longitudinal measurement was made.

Study Group

The study group consisted of 47 teacher candidates receiving formation education at a state university in the Eastern Anatolia Region and one instructor teaching the course, totaling 48 people. In the study group, 82% of the teacher candidates were female, 18% were male and the average age was 23.45 years. The raters stated that they had no previous scoring experience. At the beginning of the semester, the instructor asked the teacher candidates to form groups individually or in a team, and 47 individuals formed 20 groups. These groups were formed as teams of one, two, three, and four. The groups scored both themselves and other groups, while the instructor scored all groups. All raters evaluated both each other and themselves, and the teacher rating all individuals.

Data Collection Instrument

Sata (2020b) developed an analytic rubric as a data collection tool for assessing the oral presentation skills of teacher candidates during the performance assessment process. In the development process of the measurement tool, a criterion pool was created by scanning the literature, and expert opinions were obtained to gather evidence for content validity. The expert opinions consisted of seven people in total: two people with a PhD in the Faculty of Communication, three people with a PhD in measurement and evaluation, and two people with a PhD in linguistics. The 14-criteria pilot rubric was reduced to 10 criteria in line with expert opinions and finalized. Experts stated that four criteria were inadequate and unnecessary in measuring the relevant skill. Additionally, exploratory factor analysis was conducted to provide evidence for construct validity. It was reported that the measurement tool, consisting of ten criteria, was subsumed under a single factor. To demonstrate the reliability of the measurement tool, the McDonald's ω coefficient (since the McDonald's ω coefficient produces more consistent values in congeneric measurements) was used and found to be .891. Based on these findings, it was stated that the measurements obtained from the measurement tool were reliable (Sata, 2020b).

In the present study, a reliability and validity study was conducted to gather evidence on the reliability and validity of the measurements. Confirmatory factor analysis (CFA) was conducted to provide evidence of the validity of the measurements obtained from the measurement tool. The fit indices obtained from the CFA were as follows: χ^2 /df = 9.707, CFI = .984, NNFI = .979, NFI = .982, RMSEA (%GA) = .082 (.074 - .090), and SRMR = .012. Since the rating performances of many raters over time were evaluated in the study, the inter-rater consistency was not examined, and the internal consistency coefficients were calculated. McDonald's ω and Cronbach's α coefficients were calculated to assess the reliability of the measurements and were estimated to be ω (95% CI) = .980 (.979 - .982) and α (95% CI) = .980 (.879 - .982), respectively. As a result, evidence was provided for the reliability and validity of the measurements obtained from the measurement tool.

Data Collection

As the study aimed to examine scorer variation over time, data collection from 20 groups was conducted and presented in Table 1.

Table 1. Presentation durations and times of the groups.

Time	Duration	Presenting groups
Week 1	35+28+29 minutes	1-2-3
Week 2	32+31+25 minutes	4-5-6
Week 3	27+22+26 minutes	7-8-9
Week 4	33+24+23+27 minutes	10-11-12-13
Week 5	24+28+27+31 minutes	14-15-16-17
Week 6	26+22+26 minutes	18-19-20

Upon examination of Table 1, it can be observed that at least three groups presented every week and that the presentation durations were similar. To ensure that other factors in the assessment of the presentation skills of the teacher candidates did not interfere with the measurements, it was ensured that the presentation topics of all groups were of parallel difficulty. In addition, all teacher candidates were given a one-week rater training on the criteria of the analytic rubric and what each performance level meant. In the one-week rater training given to the raters, the meaning of the criteria in the measurement tool and how to assign each score was explained through a sample presentation video. A total of 35 minutes of training was given and 20 minutes were spent on discussions on the criteria, aiming to create a common consensus.

Data Analysis

The study aimed to examine the changes over time in the performance assessment process within the scope of the research, and this was accomplished through the use of the many-facet Rasch model (MFRM), which allows the conversion of each source of variability into a single scale through logarithmic transformation, enabling their comparison with each other (Kim et al., 2012; Linacre, 1996). Additionally, the MFRM provides the ability to investigate the interactions between sources of variability (Kassim, 2007). The applicability of the many-facet Rasch model is increased by its ability to analyze both dichotomous and polytomous measurements simultaneously. Since the main effects of the sources of variability and their mutual interactions were considered in the performance evaluation process examined for changes over time, the MFRM is a valuable measurement model. In this study, standardized differences (Signed Area Index, SAI) and interaction terms obtained through many-facet Rasch analysis were utilized to examine the changes over time in the ratings provided by self, peer, and teacher raters.

Measurements at different times were estimated as separate models to determine the standardized differences. The logarithmic values estimated for each model were divided by their standard errors to obtain standardized values. In this study, as the presentations of teacher candidates were different at different times, the estimations of the raters were modeled to have an average of zero to eliminate the effect arising from this difference. Thus, the relative changes in the rigidity or leniency behaviors of the raters in the ratings performed at different times could be examined. For this purpose, the presentations of the teacher candidates (groups) were treated as non-centered in the study. The first measurement time was taken as the baseline, and the score deviations (SAI) were calculated by comparing the other measurements with the baseline time. The SAI value was standardized using the formula given below.

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

In this equation, Mc represents the severity or leniency of the rater compared to the baseline, while Mb represents the severity or leniency of the rater at the baseline. The two values in the denominator represent the squared standard errors of the two different time points for the rater's severity or leniency. If the calculated $Z_{SAI_{Diff}}$ value for two different times is outside the ±1.96 range, then there is a statistically significant difference between the two times (Raju, 1990). When the surfaces are positively oriented, positive $Z_{SAI_{Diff}}$ values indicate that the rater became more lenient over time, while negative $Z_{SAI_{Diff}}$ values indicate that the rater became stricter over time (Borkan, 2017).

Another index used to determine changes over time is the interaction term, which is the common variation among the variable sources (Wolfe et al., 2007). In this approach, the changes over time of the scorers are determined by examining the interactions between the time variable included as a dummy variable in the model and the scorer surface (Borkan, 2017).

The MFRM used in the study has assumptions that need to be tested, including unidimensionality, local independence, and model-data fit (Farrokhi et al., 2012). To confirm the unidimensionality assumption, a DFA analysis was performed, and confirmed that the rubric criteria were grouped under a single factor. As unidimensionality implies local independence (Hambleton et al., 1991), this assumption was also considered to be met. Finally, standardized residual values were examined to test whether the model-data fit was achieved. It was stated that the number of standardized residuals beyond the ±2 range should not exceed 5% of the total number of observations. The number of standardized residuals beyond the ±3 range should not exceed 1% of the total data (Linacre, 2017). The total number of observations in the study was 20 groups x 48 raters x 10 criteria = 9,600, but due to missing data in 871 (9.07%) observations, the total valid number of observations was considered as 8,729. According to the findings, the number of standardized residuals beyond the ±2 range was 74 (0.85%), and the number of standardized residuals beyond the ±3 range was 42 (0.48%), indicating that the model-data fit was achieved.

Results

Initially, the study examined the changes in peer raters' ratings over time. The change in ratings for each subsequent day was analyzed in relation to the ratings on the first day. In this context, ZSAIDiff and SAIDiff values were obtained by subtracting the logit values estimated separately for each model from the logit values in the first model. These values were presented in Table 2.

Table 2. Changes over time in peer raters' ratings.

	$Z_{SAIDiff}$					SAI_{Diff}				
Rater	2-1	3-1	4-1	5-1	6-1	2-1	3-1	4-1	5-1	6-1
2	-0.16	2.56				-0.06	4.73			
6	1.77	-1.47	-2.50	-0.60	-0.90	0.64	-0.49	-0.78	-0.19	-0.30
7	2.46	0.50	-2.02	-1.63		0.91	0.25	-0.63	-0.53	
8	0.94	4.18	3.02			0.30	1.52	0.96		
9	1.81	2.63	-0.76	0.59	3.23	0.55	0.84	-0.22	0.17	1.03
12	-1.17	2.11	0.58	-3.10	-0.77	-0.39	0.95	0.21	-0.97	-0.26
13	-0.69	0.09	0.51	1.62	2.63	-0.23	0.03	0.17	0.54	0.99
14	1.02	3.27	3.05	3.50	1.63	0.34	1.09	0.95	1.09	0.52
16	0.00	-0.03	2.87	1.03	0.30	0.00	-0.01	1.12	0.35	0.11
17	1.95	-1.30	0.99	3.06	2.05	0.72	-0.44	0.33	1.04	0.74
20	1.46	-0.59	2.26	-2.78	-0.25	0.56	-0.21	0.83	-0.89	-0.09
22	-1.55	0.00	1.73	0.46	0.48	-0.55	0.00	0.66	0.16	0.18
23	-0.49	2.48	4.31			-0.16	0.90	2.14		
24	1.44	1.93	2.29	-0.13	-1.10	0.49	0.67	0.76	-0.04	-0.35
25	2.51	0.97	0.79	3.47	0.87	0.87	0.33	0.25	1.13	0.29
26	1.28	0.50	1.75	2.31	1.55	0.48	0.18	0.63	0.80	0.57
27	0.48	1.83	2.35	1.92	0.57	0.16	0.65	0.78	0.61	0.19
31	0.64	0.41	1.77	3.12	2.88	0.20	0.13	0.55	0.97	0.96
36	1.55	-0.26	1.69	2.42	1.86	0.47	-0.08	0.54	0.72	0.58
37	2.76	1.73	3.43	1.38		1.00	0.60	1.19	0.47	
39	0.74	0.56	-0.59	2.79	1.09	0.24	0.17	-0.17	0.81	0.33
42	4.68	2.76	4.64	2.73	5.43	1.78	0.92	1.58	0.83	2.02
43	3.62	1.63	0.07	0.14	0.26	1.21	0.52	0.02	0.04	0.08
45	1.38	0.41	2.15	0.21	-1.55	0.46	0.18	0.64	0.06	-0.46
46	1.14	4.45	5.39	8.02	6.13	0.33	1.42	1.64	2.62	2.05
48	-1.48	1.44	2.86	2.00	4.26	-0.44	0.45	0.87	0.58	1.49
Mean	0.56	0.74	0.96	0.95	0.88	0.23	0.31	0.34	0.32	0.31
SD	1.45	1.44	1.79	1.87	1.72	0.60	0.87	0.65	0.60	0.60

* Only raters who showed a statistically significant rater drift are presented.

When Table 2 is examined, it can be seen that the teacher candidates made statistically significant rater shifts at the group level in all scoring over time. The average score of 48 raters (self, peer, and teacher) increased by 0.56 points from the first to the second week of the study. It is observed that this increase continued to rise each week and only slightly decreased in the last week compared to the previous week. It has been stated that $Z_{SAIdiff}$ between two-time measurements should be 0.50 or higher for significant rater consistency or generosity at the group level (Swaminathan & Rogers, 1990). When the mean of the other five-time measurements is examined according to the base time, it was found that this difference was above the threshold, and raters showed a progressively generous rater shift over time. After determining the significant rater shift at the group level, the rater shift at the individual level was examined. Raters with at least one statistically significant rater shift in their scores are seen in Table 2. Rater number one is a teacher rater and did not show any rater shift over time, as they had a similar level of consistency and generosity throughout all weeks. When the logit values of the first rater for each week are examined, it can be seen that they received similar values (See Appendix 1). It is observed that some raters in Table 2 do not have different values, indicating that they did not participate in all assessments. For example, rater number two participated in the first three assessments but not the next three. When Table 2 is examined, it can be seen that 26 out of 48 raters became either more consistent or generous over time. It is observed that the majority of statistically significant rater shifts had a positive value, indicating that the raters became progressively more generous over time. In addition, some raters (numbers 7, 12, and 20) made consistent and lenient ratings over time.

In addition to standard deviations, common effects are also examined when determining rater drift, and rater drift is investigated through the interaction term. Accordingly, the time

variable was included in the model in the study, and rater*time interactions were examined, and the findings obtained are presented in Table 3.

Table 3. Rater*time interactions.

	I_{Diff}				
Rater	2-1	3-1	4-1	5-1	6-1
2	0.14	2.67			
6	2.03	-1.23	0.53		1.10
7	2.65	0.53	-1.44		
8	1.09	4.05	1.50		
9	1.92	2.86	0.24		-0.26
12	-0.89	2.11	0.42	0.10	-1.23
14	1.19	3.29	4.25	2.59	1.73
17	2.18	-1.10	1.32	0.88	1.57
18	0.66	1.23	3.17	1.98	
19	1.71	0.55	2.23	-0.31	-0.06
21	1.09	1.92	3.51	-2.47	-0.40
22	-1.27	0.06	1.86	-1.98	0.61
23	-0.28	2.42	1.21	1.32	
25	2.70	1.10	2.73	2.83	2.15
26	1.57	0.58	2.62	1.71	0.93
27	0.74	1.88	2.04	1.38	0.33
29	-1.38	1.30	0.03	-1.05	-0.09
31	0.82	0.62	0.03	2.66	1.66
36	1.61	1.19	2.19	2.08	1.09
37	2.95	0.77	1.41	1.01	
38	2.13	0.92	2.14	0.93	-1.04
39	0.81	1.18	1.50	2.64	2.81
42	4.83	2.88	5.61	2.28	3.25
43	3.71	0.91	2.36	0.07	2.98
45	1.43	1.82	2.30	0.26	0.95
46	1.13	4.20	6.15	7.41	6.15
48	-1.28	1.75	3.95	1.69	2.24
Mean	0.80	0.90	1.51	0.81	0.82
SD	1.41	1.36	1.61	1.67	1.42

Fixed (all = 0) chi-square: 1262.00 df.: 373 significance (probability): .000
Variance explained by interaction (%) : 8.94

* Only raters who showed a statistically significant rater drift are presented.

When Table 3 is examined, it is determined that the interaction between scorer and time is statistically significant at the group level (χ^2 (df) = 1260.00 (373); $p < .05$). This result is consistent with standardized differences. When both standardized differences and interaction analyses are examined, it is observed that raters who exhibit shifts over time are the same. When Table 3 is examined, it is seen that the highest rater drift is in the fourth week ($\bar{X}$= 1.51). The other four weeks have similar averages. Furthermore, the most significant rater drifts have a positive value, indicating that scorers become more lenient over time. The fact that none of the scoring by one particular rater was biased indicates that there was no statistical change in leniency or severity over time. When interaction analyses are examined, it is found that 27 out of 48 raters showed rater drift in at least one scoring. As a result, it is determined that interaction analysis and standardized differences yield similar results.

After providing evidence for scorer shifts through standardized differences and interaction analysis, rater*time interactions were graphically obtained and presented in Figure 1.

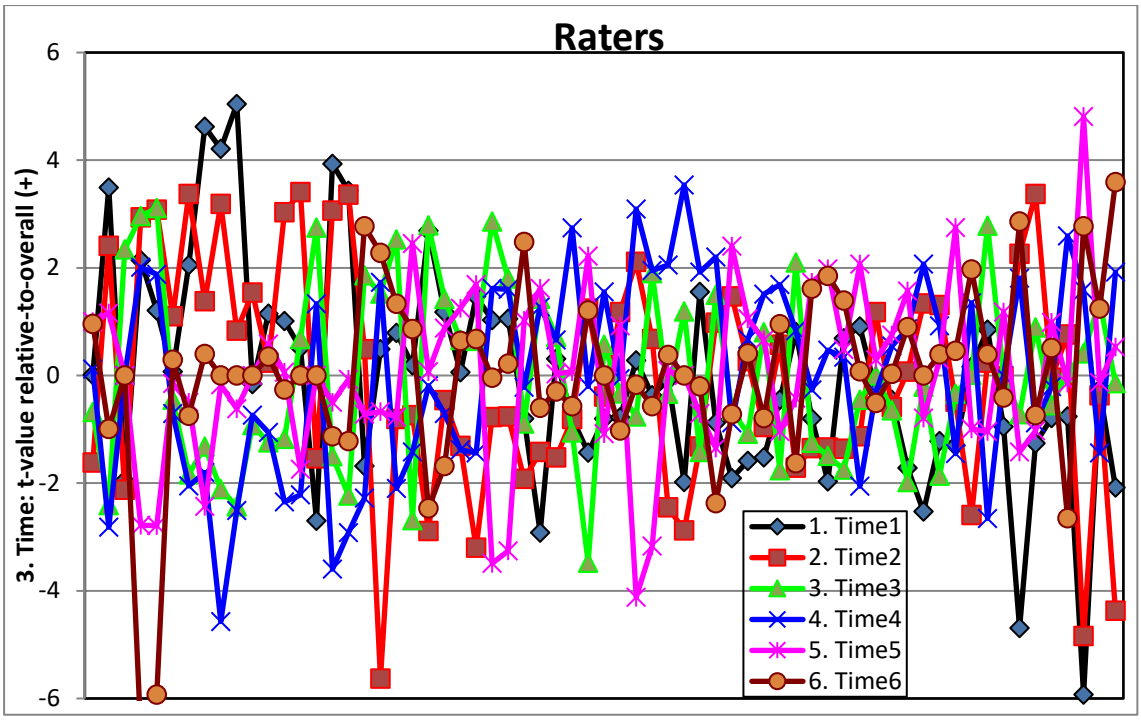

Figure 1. t-values for rater*time interactions.

When Figure 1 is examined, it is observed that some of the t-values for the scores given by the raters are greater than ±1.96, indicating that they are biased. In other words, the scores are particularly leniency at the positive end.

Discussion, Conclusion and Recommendations

This study investigated the change in raters' behaviors over time during the evaluation of teacher candidates' presentation skills. In the study, 47 teacher candidates evaluated both themselves and their peers, and a teacher evaluated all of the teacher candidates. Since the focus of the study was on the rater drift over time, analysis reports were not provided for each surface within the study, only logit values and standard errors of each measurement were reported for each week (See Appendix 1 and Appendix 2). When the study findings were examined, it was determined that the rater behaviors of the teacher candidates had statistically significant differences over time at the group level. Rater drift was observed in 26 out of 48 peer raters (54.17%). Most of the rater drift over time was positive, meaning that raters became more lenient. Of the 26 raters exhibiting rater drift, 19 (73.07%) became more leniency over time, 4 (15.39%) became severity over time, and 3 (11.54%) gave both severity and more leniency ratings over time. Similar results were found in a study conducted by Borkan (2017), tended to give more leniency ratings over time. When the literature was examined, it was seen that there were studies that supported the study's results and those that did not. For example, in the study conducted by Congdon and McQueen (2000), rater severity (the rater's severity compared to other raters on the same day) was observed to change from day to day without a predictable pattern, and 10 raters significantly deviated from their initial estimates towards the end, with 9 raters becoming severity and 1 rater becoming more leniency. Some researchers have also stated that rater drift results are meaningless (Leckie & Baird, 2011).

Another finding reached within the scope of the study is that teacher assessments did not

exhibit any significant rater drift over the six-week period, indicating consistent scoring patterns over time. Considering the difficulty in achieving objectivity in the performance evaluation process and the fact that irrelevant variance is often mixed into measurements (Mesick, 1995), it is important to use teacher evaluations in addition to self and peer assessments to reduce the limitations of the latter methods. This way, the various self and peer assessment practice limitations can be mitigated.

In the process of performance evaluation, the participation of a large number of raters increases the reliability and validity of measurements (Karakaya, 2015). Therefore, many raters have also been used in the present study. As revealed in the study, there are statistical differences in the scoring behavior of many raters over time, and it is observed that not taking the scores of these raters into account in measurement situations where important decisions are made would contribute to the reliability and validity of measurements. It was found that some raters made both severity and leniency assessments over time, increasing the scoring inconsistency. Furthermore, when the logit values of the raters were examined (Appendix 1 and Appendix 2), it was seen that the teacher rater was consistently a severity rater throughout all times, but this was because self and peer raters drift towards leniency over time. Particularly, peer raters' preference for the rubric's midpoint and the higher end of each criterion caused the teacher rater to appear severe. This is due to the tendency of self and peer raters to give themselves higher scores and the fear of failure (Uzun & Yurdabakan, 2011). Personal relationships influence the peer rater who drifts towards severity scoring over time and punishes peers with whom they have a poor relationship (Alici, 2010). Although rater training and feedback were given to peer raters, it was found that rater drift did not decrease (Harik et al., 2009). Therefore, it is necessary to consider the evaluations of teacher raters as well, as using only self and peer raters as decision-makers in performance evaluation would negatively affect the reliability and validity of measurements. Moreover, keeping the number of raters constantly high is important, as the reliability and validity of measurements will increase accordingly.

This study examined the change in ratings over time in the assessment of teacher candidates' oral presentation skills. The results made an important contribution to the literature. However, there are also some limitations. First, although the general framework of the oral presentations in the study was the same, there were differences in the content. Second, the fact that the majority of the peer raters were female made it difficult to assess performance by gender. Third, the fact that the raters were inexperienced and did not receive any rater training constituted an important constraint in terms of the reliability of the measurements. Fourth, some groups were evaluated by fewer people because the study was conducted in a limited period of time and not all teacher candidates participated every week. It is believed that taking these limitations into account in future studies on rater drift will contribute to the reliability and validity of the measurements obtained in the study.

Based on the results obtained within the scope of the research, some suggestions were presented for future rater drift studies and researchers working in this field. In this study, only the results of the evaluations made by novice raters are included. In future studies, the scoring behaviors of expert and novice raters over time can be examined. Thus, the change in scoring experience over time can be examined. Since the majority of the raters in this study were women, it was found that differential rater leniency was more common. Confirming this effect of gender on scoring in future studies may provide evidence for the reliability and validity of the measurements. In this study, oral presentation skills were considered and the main effects of raters over time were examined. In future studies, the results of the study can be confirmed by examining the effect of time in the process of evaluating higher level skills

such as writing and experimentation. In education and psychology, in addition to measuring unidimensional constructs, multidimensional constructs are also measured. In this context, a unidimensional construct was taken into account in the current study and multidimensional constructs were not examined. It is recommended that future studies and researchers working in this field should examine the role of time in the evaluation of multidimensional constructs. The availability of statistical package programs used in the measurement of multidimensional constructs makes this possible (Koyuncu & Sata 2023).

References

Alaz, A., & Yarar, S. (2009, May). *Classroom teachers' preferences and reasons in the measurement and evaluation process*. I. International Education Research Congress. Canakkale Onsekiz Mart University, Canakkale.

Alici, D. (2010). Other measurement tools and methods used in evaluating student performance. In Tekindal S. (Ed.), *Measurement and evaluation in education* (pp. 127-168). Pegem Akademi Publishing.

Ananiadou, K., & Claro, M. (2009). 21St century skills and competences for new millennium learners in OECD countries. *OECD education working papers*, *41*, OECD Publishing.

Arik, R. S., & Kutlu, O. (2013). Scaling the competency of teachers' measurement and evaluation field based on judge decisions. *Journal of educational sciences research*, *3*(2), 163-196. https://doi.org/10.12973/jesr.2013.3210a

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, policy & practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Board of Education (2005). *Introduction booklet of primary school grades 1-5 curriculum*. Ministry of National Education.

Borkan, 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

Boud, D. (2013). *Enhancing learning through self-assessment.* Routledge. https://doi.org/10.4324/9781315041520

Case, H. (1997). An examination of variation in rater severity over time: A study in rater drift. Objective measurement: *Theory into practice*, *5*, 1-38.

Cepni, S. (2010). *Introduction to research and project work*. Celepler Publishing.

Colvin, S., & Vos, E. K. (1997). Authentic assessment models for statistics education. *The assessment challenge in statistics education*, 27-36.

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

Dikli, S. (2003). Assessment at a distance: Traditional vs. alternative assessments. *Turkish online journal of educational technology-TOJET*, *2*(3), 13-19.

Dishon, G., & Gilead, T. (2021). Adaptability and its discontents: 21st-century skills and the preparation for an unpredictable future. B*ritish journal of educational studies, 69*(4), 393-413. https://doi.org/10.1080/00071005.2020.1829545

Dogan, C. D., & Uluman, M. (2017). A Comparison of rubrics and graded category rating scales with various methods regarding raters' reliability. *Educational sciences: Theory and practice*, *17*(2), 631-651. https://doi.org/10.12738/estp.2017.2.0321

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. https://doi.org/10.4236/ce.2013.46A005

Duban, N., & Kucukyilmaz, E. A. (2008). Classroom teacher candidates' views on the use of alternative assessment techniques in application schools. *Elementary education online, 7*(3), 769-784.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical assessment, research, and evaluation*, *14*(1), 1-11. https://doi.org/10.7275/jg4h-rb87

Engelhard Jr, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model*. ETS Research Report Series, *2003*(1), i-60. https://doi.org/10.1002/j.2333-8504.2003.tb01893.x

Erman-Aslanoglu, A. & Sata, 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

Erman-Aslanoglu, A. (2017). Evaluation of an individual within a group: Peer and self-assessment. *Bogazici university journal of education, 34*(2), 35-50.

Erman-Aslanoglu, A. (2022). Examining the effects of peer and self-assessment practices on writing skills. *International journal of assessment tools in education, 9*(Special Issue), 179-196. https://doi.org/10.21449/ijate.1127815

Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Innovations in Education and training International*, *32*(2), 175-187. https://doi.org/10.1080/1355800950320212

Farrokhi, F., Esfandiari, R., & 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*(11), 70-77.

Farrokhi. F., Esfandiari. R., & Schaefer. E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT journal, 34*(1). 79-101. https://doi.org/10.37546/JALTJJ34.1-3

Gelbal, S., & Kelecioglu, H. (2007). Teacher competency perceptions and problems encountered in measurement and evaluation methods. *Hacettepe university journal of education,* (33), 135-145.

Gocer, A., Arslan, S., & Cayli, C. (2017). Process-oriented complementary assessment tools and methods for determining student development in Turkish education. *Suleyman Demirel university journal of social sciences institute,* (28), 263-292.

Gomleksiz, M. N., Yetkiner, A., & Yildirim, F. (2011). Teachers' views on the use of alternative assessment and evaluation techniques in life studies class. *Education sciences, 6*(1), 823-840.

Guler, N. (2012). *Measurement and assessment in education*. Pegem Akademi Publishing. https://doi.org/10.14527/9786053641247

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating. *Int. J. Sci. Educ., 25*(12), 1509-1528. https://doi.org/10.1080/0950069022000038268

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items.* Routledge. https://doi.org/10.4324/9780203850381

Hamayan, E. V. (1995). Approaches to alternative assessment. *Annual review of applied linguistics, 15*, 212-226. https://doi.org/10.1017/S0267190500002695

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

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

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the golden sate examination. *Journal of educational measurement, 38*(2), 121-145. https://doi.org/10.1111/j.1745-3984.2001.tb01119.x

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it?. *Psychological methods, 5*(1), 64–86. https://doi.org/10.1037/1082-989X.5.1.64

Karakaya, I. (2015). Comparison of self, peer and instructor assessments in the portfolio assessment by using many facet Rasch model. *Journal of education and human development, 4*(2), 182-192. https://doi.org/10.15640/jehd.v4n2a22

Kassim. A.N.L. (2007, June). *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.

Kilic, D., & Gunes, P. (2016). Self, peer, and teacher assessment with grading rubrics. *Mehmet Akif Ersoy university journal of education faculty*, *1*(39), 58-69. https://doi.org/10.21764/efd.93792

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

Kooken, J., Welsh, M. E., McCoach, D. B., Miller, F. G., Chafouleas, S. M., Riley-Tillman, T. C., & Fabiano, G. (2017). Test order in teacher-rated behavior assessments: Is counterbalancing necessary?. *Psychological assessment, 29*(1), 98-109. https://doi.org/10.1037/pas0000314

Kostereliogl u, İ., & Celen, Ü. (2016). Evaluation of the effectiveness of self-assessment method. *Ilkogretim online*, *15*(2), 671-681. https://doi.org/10.17051/io.2016.44304

Koyuncu, M. S. & Sata, M. (2023). Using ACER ConQuest program to examine multidimensional and many-facet models. *International journal of assessment tools in education, 10*(2), 279-302. https://doi.org/10.21449/ijate.1238248

Kutlu, O., Dogan, C.D., & Karakaya, I., (2010). *Determination of student achievement: Performance-based and portfolio-based authentic assessment and evaluation practices.* Pegem Akademi Publishing.

Lamprianou, I. (2006). The stability of marker characteristics across tests of the same subject and across subjects. *Journal of applied measurement*, *7*(2), 192-205.

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. (1996). Generalizability theory and many-facet Rasch measurement. *Objective measurement: Theory into practice*, *3*, 85-98.

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

Linn, R. L. (2008). *Measurement and assessment in teaching*. Pearson Education

Maier, A., Adams, J., Burns, D., Kaul, M., Saunders, M., & Thompson, C. (2020). *Using performance assessments to support student learning: how district ınitiatives can make a difference. performance assessment case study series.* Learning policy institute. i-68. Palo Alto. https://doi.org/10.54300/213.365

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

McNamara, T. F., & Adams, R. J. (1991). Exploring rater characteristics with Rasch techniques. In Selected papers of the 13th Language Testing Research Colloquium (LTRC). Educational Testing Service, International Testing and Training Program Office.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Modarresi, G., Jalilzadeh, K., Coombe, C., & Nooshab, A. (2021). Validating a test to measure translation teachers' assessment literacy. *Journal of Asia TEFL*, *18*(4), 1503-1511. https://doi.org/10.18823/asiatefl.2021.18.4.31.1503

Mulqueen, C., Baker, D., & Dismukes, R. K. (2000, April). *Using multifacet Rasch analysis to examine the effectiveness of rater training.* In 15th Annual Conference for the Society for Industrial and Organizational Psychology. https://doi.org/10.1037/e540522012-001

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

Nalbantoglu Yilmaz, F. (2017). Analysis of the rater effects on the scoring of diagnostic trees prepared by teacher candidates with the many-facet Rasch model. *Online submission*, *8*(18), 174-184. https://doi.org/10.15345/iojes.2016.02.020

National Research Council. (2001). *Classroom assessment and the national science education standards.* National Academies Press.

Noonan, B., & Duncan, C. R. (2005). Peer and self-assessment in high schools. *Practical assessment, research, and evaluation*, *10*(1), 1-8. https://doi.org/10.7275/a166-vm41

Oren, F. S., Ormanci, U., & Evrekli, E. (2014). The alternative assessment-evaluation approaches preferred by pre-service teachers and their self-efficacy towards these approaches. *Educational sciences: Theory & practice*, *11*(3), 1690-1698.

Orlova, N. (2019). Student peer performance evaluation: importance of implementation for group work enhancement. *Science and education a new dimension: Pedagogy and psychology*, 26-29. https://doi.org/10.31174/SEND-PP2019-207VII84-05

Ozpinar, I. (2021). Self, peer, group, and instructor assessment: A glimpse through the window of teacher competencies. *Cumhuriyet international journal of education, 10*(3), 949-973. https://doi.org/10.30703/cije.754885

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical assessment, research, and evaluation*, *13*(4), 1-11. https://doi.org/10.7275/0qpc-ws45

Park, Y. S. (2011). *Rater drift in constructed response scoring via latent class signal detection theory and item response theory.* Columbia University. https://doi.org/10.7916/D8445TGR

Petra, T. Z. H. T., & Ab Aziz, M. J. (2020, April). Investigating reliability and validity of student performance assessment in higher education using Rasch model. In Journal of Physics: Conference Series 1529(4), 042088. *IOP Publishing.* https://doi.org/10.1088/1742-6596/1529/4/042088

Quellmalz, E. (1980). *Problems in stabilizing the judgment process* (Vol. CSE Report No. 136). Center for the Study of Evaluation.

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://doi.org/10.1177/014662169001400208

Raymond, M. R., Harik, P., & Clauser, B. E. (2011). The impact of statistically adjusting for rater effects on conditional standard errors of performance ratings. *Applied Psychological measurement, 35*(3), 235-246. https://doi.org/10.1177/0146621610390675

Rennert-Ariev, P. (2005). A theoretical model for the authentic assessment of teaching. *Practical assessment, research, and evaluation, 10*(2), 1-12. https://doi.org/10.7275/a7h7-4111

Sad, S. N., & Goktas, O. (2013). Examination of traditional and contemporary measurement and evaluation approaches of academic staff. *Ege education journal*, *14*(2), 79-105.

Sata, M. & Karakaya, I. (2022). Investigating the impact of rater training on rater errors in the process of assessing writing skill. *International journal of assessment tools in education*, *9*(2), 492-514. https://doi.org/10.21449/ijate.877035

Sata, M. (2020a). Quantitative research approaches. In E. Oğuz (Ed.), *Research methods in education* (pp. 77-98). Egiten Kitap Publications.

Sata, M. (2020b, November). *Evaluation of university students' oral presentation skills by their peers.* 13th International Education Community Symposium. Online. Turkey.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher*, *29*(7), 4-14. https://doi.org/10.3102/0013189X029007004

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

Szőkol, I., Szarka, K., & Hargaš, J. (2022). The functions of educational evaluation. *R&Amp;E-SOURCE*, (S24). https://doi.org/10.53349/resource.2022.iS24.a1112

Tunkler, V. (2019). Investigation of the contribution of peer assessment to pre-service teachers' professional knowledge and skills. *Marmara university Atatürk education faculty journal of educational sciences, 50*(50), 206-221. https://doi.org/10.15285/maruaebd.525171

Uto, M. (2022). A Bayesian many-facet Rasch model with Markov modeling for rater severity drift. *Behavior research methods*, 1-19. https://doi.org/10.3758/s13428-022-01997-z

Uzun, A., & Yurdabakan, I. (2011). An investigation of elementary school students' attitudes towards self-assessment. *Mehmet Akif Ersoy university journal of education faculty, 11*(22), 51-69.

Wayda, V., & Lund, J. (2005). Assessing dispositions: An unresolved challenge in teacher education. *Journal of physical education, recreation & dance*, *76*(1), 34-41. https://doi.org/10.1080/07303084.2005.10607317

Wesolowski, B. C., Wind, S. A., & Engelhard Jr, G. (2017). Evaluating differential rater functioning over time in the context of solo music performance assessment. *Bulletin of the council for research in music education*, (212), 75-98. https://doi.org/10.5406/bulcouresmusedu.212.0075

Wigglesworth, G. (1994). Patterns of rater behaviour in the assessment of an oral interaction test. *Australian review of applied linguistics*, *17*(2), 77-103. https://doi.org/10.1075/aral.17.2.04wig

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (1999, April). *Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model.* Annual Meeting of the American Educational Research Association. Montreal, Quebec, Canada.

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.* Research Report No. 2007-2. College Board.

Yildiz, S. (2018). Developing a self-assessment scale for fractions. *Mustafa Kemal university journal of faculty of education*, *2*(3), 30-44.

Yurdabakan, I. (2012). The effect of peer and collaborative assessment training on pre-service teachers' self-assessment skills. *Education and science*, *37*(163), 190-202.

Zhu, W., & Cole, E. L. (1996). Many-faceted Rasch calibration of a gross motor instrument. *Research quarterly for exercise and sport*, *67*(1), 24-34. https://doi.org/10.1080/02701367.1996.10607922

Appendix 1. Logit values for the six-week scores of raters.

	Week 1		Week 2		Week 3		Week 4		Week 5		Week 6	
Rater	Logit	SE	Logit	SE	Logit	SE	Logit	SE	Logit	SE	Logit	SE
1	-1.23	0.21	-1.45	0.20	-1.44	0.22	-1.36	0.20	-1.12	0.19	-1.07	0.21
2	0.59	0.26	0.53	0.26	5.32	1.83						
3	0.52	0.26	0.83	0.29	0.45	0.26	0.36	0.23	0.58	0.23	0.63	0.26
4	0.66	0.26	1.20	0.32	0.52	0.27	0.38	0.46				
5	-0.08	0.23	-0.24	0.23	-0.63	0.23	-0.63	0.21	-0.20	0.21	-0.27	0.23
6	0.03	0.24	0.67	0.27	-0.46	0.23	-0.75	0.20	-0.16	0.21	-0.27	0.23
7	-0.08	0.23	0.83	0.29	0.17	0.44	-0.71	0.21	-0.61	0.23		
8	-0.54	0.22	-0.24	0.23	0.98	0.29	0.42	0.23				
9	-1.14	0.21	-0.59	0.22	-0.30	0.24	-1.36	0.20	-0.97	0.20	-0.11	0.24
10	0.95	0.37	0.36	0.31	1.51	0.63	0.33	0.27	0.85	0.23	0.12	0.24
11	0.03	0.24	-0.24	0.23	0.28	0.31	-0.41	0.29	0.44	0.25	0.24	0.30
12	0.15	0.24	-0.24	0.23	1.10	0.38	0.36	0.27	-0.82	0.20	-0.11	0.24
13	0.09	0.24	-0.14	0.23	0.12	0.25	0.26	0.23	0.63	0.23	1.08	0.29
14	-0.85	0.22	-0.51	0.25	0.24	0.25	0.10	0.22	0.24	0.22	-0.33	0.23
15	0.96	0.28	0.46	0.26	1.17	0.31	1.10	0.27	1.01	0.24	0.84	0.27
16	0.39	0.25	0.39	0.26	0.38	0.26	1.51	0.30	0.74	0.23	0.50	0.26
17	0.03	0.24	0.75	0.28	-0.41	0.24	0.36	0.23	1.07	0.24	0.77	0.27
18	-0.39	0.22	-0.24	0.23	-0.01	0.25	0.15	0.22	-0.45	0.23		
19	0.21	0.24	0.75	0.28	0.38	0.26	0.42	0.23	0.74	0.23	0.18	0.25
20	0.27	0.25	0.83	0.29	0.06	0.25	1.10	0.27	-0.62	0.20	0.18	0.25
21	0.03	0.24	0.33	0.25	0.74	0.28	0.65	0.24	-0.50	0.20	0.00	0.24
22	0.52	0.26	-0.03	0.24	0.52	0.27	1.18	0.28	0.68	0.23	0.70	0.27
23	-0.08	0.23	-0.24	0.23	0.82	0.28	2.06	0.44				
24	-0.29	0.23	0.20	0.25	0.38	0.26	0.47	0.24	-0.33	0.20	-0.64	0.22
25	-0.34	0.23	0.53	0.26	-0.01	0.25	-0.09	0.22	0.79	0.23	-0.05	0.24
26	0.27	0.25	0.75	0.28	0.45	0.26	0.90	0.26	1.07	0.24	0.84	0.27
27	-0.13	0.23	0.03	0.24	0.52	0.27	0.65	0.24	0.48	0.22	0.06	0.24
28	-0.02	0.24	0.26	0.25	-0.30	0.24	0.53	0.24	-0.07	0.21	0.37	0.25
29	0.21	0.24	-0.34	0.22	0.66	0.27	0.21	0.23	-0.03	0.21	-0.38	0.23
30	-0.73	0.22	-0.77	0.21	-0.84	0.23	-0.67	0.21	-0.16	0.21	-0.17	0.24
31	-0.59	0.22	-0.39	0.22	-0.46	0.23	-0.04	0.22	0.38	0.22	0.37	0.25
32	0.15	0.24	-0.29	0.23	-0.41	0.24	0.05	0.22	0.15	0.21	0.37	0.25
33	1.86	0.37	1.06	0.37	1.44	0.34	1.03	0.27	2.47	0.36	1.64	0.33
34	1.74	0.35	3.99	1.84			1.81	0.33	2.14	0.37	1.75	0.34
35	1.74	0.35	1.60	0.44	1.72	0.37	2.19	0.38	2.35	0.34	2.00	0.37
36	-1.01	0.21	-0.54	0.22	-1.09	0.22	-0.47	0.24	-0.29	0.21	-0.43	0.23
37	-0.29	0.23	0.71	0.28	0.31	0.26	0.90	0.26	0.18	0.25		
38	-0.39	0.22	0.26	0.25	-0.52	0.23	0.10	0.22	0.01	0.21	0.00	0.24
39	-1.31	0.21	-1.07	0.25	-1.14	0.22	-1.48	0.20	-0.50	0.20	-0.98	0.22
40	-0.08	0.23	-0.63	0.22	-0.07	0.24	0.21	0.23	-0.29	0.21	0.43	0.25
41	-1.27	0.21	-0.96	0.36					-0.89	0.20	-1.21	0.21
42	-0.68	0.22	1.10	0.31	0.24	0.25	0.90	0.26	0.15	0.21	1.34	0.30
43	-0.82	0.21	0.39	0.26	-0.30	0.24	-0.80	0.20	-0.78	0.20	-0.74	0.22
44	0.33	0.25	0.53	0.26	0.45	0.26	0.44	0.50	0.90	0.24	0.84	0.47
45	-1.18	0.21	-0.72	0.26	-1.00	0.39	-0.54	0.21	-1.12	0.19	-1.64	0.21
46	-1.49	0.21	-1.16	0.20	-0.07	0.24	0.15	0.22	1.13	0.25	0.56	0.26
47	0.21	0.24	0.20	0.25	0.74	0.28	-0.04	0.22	0.34	0.22	0.74	0.33
48	-0.91	0.21	-1.35	0.21	-0.46	0.23	-0.04	0.22	-0.33	0.20	0.58	0.28
Mean	-0.08		0.15		0.26		0.26		0.21		0.21	
SD	0.77		0.90		1.04		0.83		0.86		0.78	

Appendix 2. Bias values for rater*time interactions.

	Week 1		Week 2		Week 3		Week 4		Week 5		Week 6	
Rater	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
1												
2	-0.53	0.25	-0.48	0.26	3.92	1.65						
3	-0.02	0.25	0.40	0.29	-0.06	0.25					0.00	0.25
4	-0.08	0.26	0.57	0.32	-0.18	0.26						
5	0.22	0.23	0.14	0.23	-0.21	0.21	-0.13	0.19			0.00	0.22
6	0.19	0.23	0.91	0.27	-0.20	0.22	-0.37	0.19			-0.13	0.22
7	0.08	0.23	1.06	0.29	0.33	0.41	-0.35	0.19				
8	-0.61	0.21	-0.27	0.23	0.84	0.29	0.36	0.22				
9	-0.38	0.20	0.19	0.22	0.47	0.22	-0.34	0.18			0.54	0.22
10	0.38	0.36	-0.06	0.31	0.94	0.62	-0.11	0.26	0.40	0.20	-0.46	0.23
11	-0.02	0.23	-0.21	0.23	0.25	0.30	-0.30	0.27	0.10	0.22	0.11	0.28
12	0.20	0.23	-0.09	0.23	1.12	0.37	0.48	0.25	0.23	0.19	-0.08	0.22
13	-0.22	0.23	-0.36	0.23	-0.16	0.24	0.02	0.21	-0.67	0.18	0.60	0.27
14	-0.65	0.21	-0.26	0.25	0.40	0.24	0.33	0.21	0.12	0.21	-0.20	0.22
15	0.05	0.28	-0.31	0.26	0.27	0.30	0.24	0.26	0.23	0.20	-0.15	0.26
16	-0.23	0.24	-0.12	0.26	-0.21	0.25	0.87	0.29	-0.09	0.23	-0.21	0.24
17	-0.36	0.23	0.43	0.28	-0.71	0.22	0.03	0.22	-0.08	0.22	0.23	0.26
18	-0.21	0.22	0.00	0.23	0.19	0.24	0.39	0.21	0.42	0.23		
19	-0.22	0.24	0.41	0.28	-0.03	0.25	0.06	0.22	-0.32	0.21	-0.31	0.23
20	0.03	0.24	0.69	0.29	-0.13	0.23	0.87	0.26	0.10	0.22	-0.11	0.23
21	-0.12	0.23	0.25	0.25	0.56	0.27	0.52	0.23	-0.84	0.18	-0.20	0.22
22	-0.06	0.25	-0.49	0.23	-0.04	0.26	0.61	0.26	-0.67	0.18	0.02	0.25
23	-0.48	0.23	-0.57	0.23	0.38	0.27	1.59	0.43	-0.07	0.21		
24	-0.23	0.22	0.32	0.24	0.42	0.25	0.56	0.22	-0.34	0.19	-0.57	0.21
25	-0.45	0.22	0.47	0.26	-0.10	0.23	-0.11	0.20	0.43	0.22	-0.24	0.22
26	-0.42	0.24	0.16	0.28	-0.22	0.25	0.23	0.24	0.15	0.23	0.03	0.26
27	-0.38	0.22	-0.14	0.24	0.26	0.26	0.42	0.23	0.04	0.21	-0.25	0.23
28	-0.14	0.23	0.23	0.25	-0.34	0.22	0.45	0.23	-0.29	0.19	0.15	0.24
29	0.15	0.24	-0.30	0.22	0.61	0.26	0.24	0.21	-0.17	0.19	-0.43	0.21
30	-0.20	0.21	-0.20	0.21	-0.21	0.21	0.02	0.19	0.24	0.19	0.28	0.22
31	-0.46	0.21	-0.21	0.22	-0.27	0.22	0.16	0.20	0.31	0.20	0.36	0.24
32	0.12	0.23	-0.22	0.22	-0.33	0.22	0.14	0.21	0.00	0.20	0.25	0.24
33	0.29	0.36	-0.34	0.37	-0.10	0.33	-0.45	0.25	0.64	0.35	-0.05	0.32
34	-0.14	0.35	2.01	1.64			-0.03	0.32	0.01	0.36	-0.25	0.33
35	-0.19	0.35	-0.17	0.44	-0.18	0.36	0.27	0.37	0.16	0.34	-0.06	0.36
36	-0.39	0.21	0.10	0.22	-0.36	0.21	0.26	0.22	0.20	0.19	0.12	0.21
37	-0.59	0.22	0.46	0.28	0.00	0.25	0.58	0.24	-0.27	0.23		
38	-0.30	0.22	0.41	0.25	-0.35	0.22	0.26	0.21	-0.03	0.19	0.01	0.22
39	-0.30	0.20	-0.04	0.25	-0.02	0.21	-0.19	0.18	0.41	0.18	0.02	0.20
40	0.00	0.23	-0.47	0.21	0.06	0.23	0.35	0.21	-0.28	0.19	0.40	0.24
41	-0.23	0.20	0.08	0.36					0.12	0.18	-0.16	0.20
42	-1.03	0.21	0.78	0.31	-0.12	0.24	0.51	0.24	-0.37	0.20	0.76	0.29
43	-0.30	0.21	0.94	0.26	0.25	0.22	-0.09	0.19	-0.28	0.18	-0.23	0.21
44	-0.23	0.24	0.07	0.26	-0.09	0.25	-0.03	0.47	0.13	0.22	0.15	0.45
45	-0.19	0.20	0.28	0.26	0.08	0.36	0.56	0.19	-0.12	0.18	-0.61	0.20
46	-1.23	0.20	-0.91	0.20	0.15	0.23	0.40	0.21	1.03	0.23	0.60	0.25
47	-0.10	0.24	-0.01	0.24	0.42	0.27	-0.22	0.20	-0.12	0.20	0.31	0.31
48	-0.47	0.21	-0.85	0.21	0.02	0.22	0.46	0.20	0.01	0.19	0.86	0.26
Mean	-0.22		0.09		0.16		0.22		0.01		0.03	
SD	0.31		0.53		0.68		0.39		0.36		0.34	