International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 492–514

https://doi.org/10.21449/ijate.877035

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Investigating the Impact of Rater Training on Rater Errors in the Process of Assessing Writing Skill

Mehmet Sata[1,*], **Ismail Karakaya**[2]

[1]Agri Ibrahim Cecen University, Faculty of Education, Department of Educational Sciences, Agri, Türkiye
[2]Gazi University, Faculty of Gazi Education, Department of Educational Sciences, Ankara, Türkiye

ARTICLE HISTORY

Received: Feb. 08, 2021
Revised: Apr. 15, 2022
Accepted: May 14, 2022

Keywords:
Rater training,
Rater errors,
Many facet Rasch model,
Validity,
Reliability.

Abstract: In the process of measuring and assessing high-level cognitive skills, interference of rater errors in measurements brings about a constant concern and low objectivity. The main purpose of this study was to investigate the impact of rater training on rater errors in the process of assessing individual performance. The study was conducted with a pretest-posttest control group quasi-experimental design. In this research, 45 raters were employed, 23 from the control group and 22 from the experimental group. As data collection tools, a writing task that was developed by IELTS and an analytical rubric that was developed to assess academic writing skills were used. As part of the experimental procedure, rater training was provided and this training was implemented by combining rater error training and frame of reference training. When the findings of the study were examined, it was found that the control and experimental groups were similar to each other before the experiment, however, after the experimental process, the study group made more valid and reliable measurements. As a result, it was investigated that the rater training given had an impact on rater errors such as rater severity, rater leniency, central tendency, and Halo effect. Based on the obtained findings, some suggestions were offered for researchers and future studies.

1. INTRODUCTION

Cognitive skills are divided into two categories: lower-order and higher-order. Lower-order cognitive skills in Bloom's Taxonomy include behaviors that belong to the remembering and understanding levels and these behaviors do not change from learner to learner, are measured by traditional tools, and are result-oriented. Higher-order cognitive skills, on the other hand, are process-oriented, are measured by complementary measurement and assessment tools (such as essay, portfolio, performance task, etc.), and their acquisition takes more time compared to lower-order cognitive skills (Kutlu et al., 2014). Kutlu et al. (2014) indicated that higher-order cognitive skills are a combination of cognitive, affective, and psychomotor characteristics of an individual when he/she displays his/her talents. Since higher-order cognitive skills are a significant indicator of the development of success, measuring them reliably and validly is of paramount importance (Haladyna, 1997).

*CONTACT: Mehmet Sata ✉ msata@agri.edu.tr Agri Ibrahim Cecen University, Faculty of Education, Department of Educational Sciences, Agri, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

It was indicated that measuring and assessing higher-order cognitive skills with traditional measurement tools is not appropriate, and complimentary measurement and assessment tools should be employed more for this purpose (Ebel, 1965; Kutlu et al., 2014). It seems more appropriate to use performance assessment to measure and assess higher-order cognitive skills consistently and accurately (Johnson et al., 2008). Performance assessment was defined as the activities that are done to determine an individual's strengths and weaknesses by observing him/her and take actions to make these better (Bennet, 1998). Performance assessment is different from traditional assessment methods due to the following characteristics: a) performance assessment is based on real-life events, b) performance assessment is process-oriented, and c) performance assessment prompts the individual to think more (Brown & Hudson, 1998; Khattri et al., 1995; Moore, 2009).

While performance assessment provides significant advantages in measuring higher-order cognitive skills, the objectivity of measurements is an important implication problem in the process of assessing an individual's performance. It is quite difficult in practice for performance assessment to be as objective as traditional assessment methods (Romagnano, 2001). When the literature is examined, it was seen that many methods were suggested and employed for the objectivity of measurements in performance assessment-based studies. These methods are automated scoring (Attali et al., 2010; Burstein et al., 1998; Landauer et al., 2003), employing more than one rater (Gronlund, 1977; Kubiszyn & Borich, 2013), using rubrics (Dunbar et al., 2006; Ebel & Frisbie, 1991; Kutlu et al., 2014; Oosterhof, 2003), and rater training (Bernardin & Buckley, 1981; Haladyna, 1997; Ilhan & Cetin, 2014; Lumley & McNamara, 1995). It was emphasized that regardless of the method used in the performance assessment process, it is often quite difficult to ensure consistency between raters and assessors (Haladyna, 1997). In other words, regardless of the method employed, there is a possibility that some external factors other than individual performance often interfere with the measurements in the process of performance assessment. These inconsistencies that occur in the performance assessment process are defined as "rater effects/behaviors / bias" (Farrokhi et al., 2011; Haladyna, 1997; Ilhan, 2015).

If one or more of the rater behaviors are involved in the performance of the individual in the performance assessment process, the amount of error in the estimations made while determining the individual's ability level will be high, so the validity of the inferences made according to these values may be low. Rater behaviors directly threaten validity because they are attributed to a variance that is unrelated to the measured structure (Abu Kassim, 2011; Brennan, Gao & Colton, 1995; Congdon & McQueen, 2000; Farrokhi et al., 2011). In this context, it is important to determine rater behaviors in the process of scoring individual performance and to bring these behaviors to a minimum or controllable level or eliminate them (Kim, 2009; Linacre, 1994).

In the performance evaluation process, one of the methods used to reduce or control rater behaviors that interfere with measurements is rater training. Rater training is widely used to reduce the variance of raters (Brijmohan, 2016). The main purpose of rater training is to explain the assessment tools to raters through sample applications and to establish a common understanding and conceptualization among raters (Fahim & Bijani, 2011). Rater training can reduce, but not eliminate, variability in rater behaviors. One of the purposes of rater training is to increase the consistency between raters and within raters by observing factors such as experience, scoring style or scoring preference, giving feedback to raters (Kim, 2009).

Many rater training patterns/models have been proposed to reduce the raters' biases, increase the accuracy of the assessment, improve observation skills, and increase behavioral accuracy and rater reliability (Woehr & Huffuct, 1994; Zedeck & Cascio, 1982). The most preferred of these patterns are; i) Self-Leadership Training (SLT), ii) Behavioral Observation Training (BOT), iii) Rater Variability Training (RVT), iv) Performance Dimension Training (PDT), v)

Rater Error Training (RET), vi) Frame-of-Reference Training (FORT). It is seen that each rater training in the literature has different approaches. Regardless of which rater training patterns are used, the main target training is expected to increase rater reliability and accuracy and decrease rater behaviors. Since this research did not examine which rater design is better, this discussion was not entered into. In this study, RET and FORT designs were combined based on the literature in order to get maximum efficiency from rater training.

Rater training methods were used in this study to determine rater behaviors and reduce them to a controllable level in the performance assessment process. The main purpose of rater training is to enable raters to develop a common-sense towards student performance and criteria of assessment preferences (Eckes, 2008; Shale, 1996). In other words, it is ensured that the assessment is done validly and reliably (Moser et al., 2016). Since the scores students get from an open-ended exam consist of both the performance of the student and the rater's interpretation of the student's performance, it creates a constant validity concern in the test results (Ellis, Johnson & Papajohn, 2002; McNamara, 1996). If the decisions that are made based on test results are vital, rater behaviors should be determined and these behaviors should be reduced to an acceptable level (Ellis et al., 2002). When the literature was examined, it was seen that many rater training designs were suggested and used (Bernardin & Buckley, 1981; Feldman et al., 2012; Haladyna, 1997; Hauenstein, & McCusker, 2017; Stamoulis & Hauenstein, 1993; Weigle, 1998; Zedeck & Cascio, 1982).

When the literature is examined, it is seen that there are many rater trainings, but the existence of such a study in the national literature has been the main motivation for conducting this study. In addition, it is thought that the relevant study is important in terms of testing the effectiveness of rater training in the evaluation of compositions. Another originality of the study is that the second language academic writing skills of Turkish students were measured for the first time with a combined rater design. In this regard, rater training was provided in this study by combining rater error training and frame of reference training designs, and its impact on rater behaviors was investigated.

For the purpose of this study, the following hypotheses were tested:

1. Before training, the raters in the experimental and control groups showed rater behaviors in the process of assessing the writing performance of students,

2. After training, the raters in the experimental group showed fewer rater behaviors than those in the control group in the process of assessing the writing performance of students.

2. METHOD

2.1. Research Design

In this study, a quasi-experimental design with control & experimental groups and pretest & posttest was employed. This pattern is a relational design because the same people are measured twice on the dependent variable. However, it is also defined as an unrelated design due to the comparison of the measurements of the experimental and control groups consisting of different participants (Howitt & Cramer, 2008). Because of these two features, pre-test post-test control group design is defined as a mixed design in the quantitative studies (Buyukozturk, 2011).

2.2. Study Group

Since there is no assumption that results obtained through Rasch models can be generalized to the universe, universe and sample were not identified in this study, instead, a study group was chosen. There were two groups involved in the study: raters and students. There were 64 raters, 12 of whom were male, and 52 of whom were female; while individuals consisted of 39 students. Both individuals/students and raters were student teachers of English at Gazi University, English Language Teaching (ELT) department. Raters were the 3rd-grade students who took the Measurement and Evaluation course, while individuals were the 1st-grade students, who took the 'Advanced Reading and Writing' course. The average age of the raters was 21.84, and they had not participated in any rater training and thus, had no experience in scoring before. The raters in the study were randomly divided into two groups (33 for the control group, and 31 for the experimental group). All the participants took place voluntarily in the research. However, 7 raters who participated in the pre-test but did not participate in the post-test were excluded from the study. Later, pre-test scores were analyzed and misfit was detected with 12 raters. These raters were also excluded from the study because the misfit negatively affected the model-data fit of the study. As a result, the study was conducted with a total number of 45 raters, 22 in the experimental group and 23 in the control group.

2.3. Data Collection Tools

A writing task (argumentative essay), personal information form, and analytical rubric were used as data collection tools in the study.

2.3.1. *The Writing task*

An argumentative essay task, which was prepared by the International English Language Testing System (IELTS) and was published as sample, was used to measure the academic writing skills (related performance) of individuals (see Appendix 1). One of the reasons for choosing this writing task is that it is authentic and reflects a real-life situation, and this provides a more valid framework for measuring student performance. Before the participants were given this task, they were informed that the researchers would not grade this task, it would be used only for academic purposes, the participation was voluntary, and they should not write their personal information on the sheets. The participants were told that they had 40 minutes to complete the task, and they are required to write an essay of within at least 250 words. The writing task was completed by 39 participants, and they were above B1 level. Later, these essays were numbered and duplicated for the rating purpose. The essays were written in the spring semester of the 2017-2018 academic year.

2.3.2. *Personal information form*

A personal information form was prepared by the researcher to collect the interests, attitudes, anxieties, and demographic information of the raters towards academic writing. A rating scale was also included in the personal information form, in which raters would write the score they gave to the essays on each criterion.

2.3.3. *Analytical rubric for academic writing skill*

To assess the essays, the researchers and a Ph.D. student from the ELT department who is knowledgeable about academic writing developed an analytical rubric for academic writing skills. While developing the rubric, a systematic process with certain steps was followed because the validity and reliability of the measurements obtained from the measurement tools developed without following a systematic process may be negatively affected. Therefore, reliability and validity should be taken into account in the process of developing rubrics (Moskal, 2000). During the development of the analytical rubric, Goodrich (1997), Haladyna (1997), Kutlu et al. (2014) and Moskal's (2000) suggestions were taken into consideration.

First of all, as the aim is to assess student teachers' academic writing skills, the purpose of the rubric was determined accordingly. In the second stage, the criteria for assessing performance (academic writing skill) were determined and sample rubrics in studies such as Weigle (2002), Hughes (2003), Brown (2004), Brown (2007), and Brookhart (2013) were examined in detail. Upon reviewing the literature, seven main criteria and 20 sub-criteria were selected and a draft form was created. Then, the draft form of the rubric was given to 11 field expertsto assess the criteria in the draft by using a measurement tool with a triple grading as (1) sufficient, (2) sufficient but should be corrected, and (3) insufficient. After the opinions of field experts were taken into account as academic writing competencies, they were presented as evidence for the content validity. For the content validity of each criterion, Lawshe's (1975) approach was taken into consideration. Since there are 11 field experts in this study, it was taken into consideration that the content validity rate (CVR) should be equal to or greater than a minimum 0.591 value in order for any criterion to have sufficient coverage in academic writing skills (Wilson, Pan & Schumsky, 2012). The CVR value for each criterion was calculated and six criteria that were less than 0.591 were removed from the draft form. Moreover, based on the feedback received from field experts, two criteria were divided into two sub-criteria. As a result, a measurement tool consisting of six main criteria and 16 sub-criteria was obtained as the final form. The final form the rubric was presented in Table 1.

Table 1. *Criteria included in the measurement of writing skill.*

Criteria		Scoring					
Main Criteria	Sub-criteria	0 score	1 score	2 score	3 score	4 score	Total score
Organization	Title of Essay	( )	( )	( )	( )	( )	( )
	Introduction-Body-Conclusion	( )	( )	( )	( )	( )	( )
	Thesis Statement	( )	( )	( )	( )	( )	( )
	Topic Sentence	( )	( )	( )	( )	( )	( )
	Supporting Sentence	( )	( )	( )	( )	( )	( )
	Appropriate Length	( )	( )	( )	( )	( )	( )
Content	Topic Relevance	( )	( )	( )	( )	( )	( )
	Idea Development	( )	( )	( )	( )	( )	( )
Coherence	Coherence	( )	( )	( )	( )	( )	( )
Cohesion	Linking	( )	( )	( )	( )	( )	( )
Grammar	Accuracy of Grammatical Forms	( )	( )	( )	( )	( )	( )
	Syntatic Complexity	( )	( )	( )	( )	( )	( )
Vocabulary	Word Choice	( )	( )	( )	( )	( )	( )
	Lexical Range	( )	( )	( )	( )	( )	( )
Mechanics	Spelling	( )	( )	( )	( )	( )	( )
	Punctuation	( )	( )	( )	( )	( )	( )

When Table 1 is examined, it can be seen that the measurement tool consists of six main and 16 sub-criteria with a five-point rating. Upon reviewing the student teachers' essays, it was seen that most of them did not give a title to their essays even though they were told to do it, so the sub-criterion of 'Title of Essay' was excluded from the study because it distorted the data structure. After the CVR value was calculated for each criterion, the content validity index (CVI) value was calculated for the measuring tool, and this value was found to be 0.750. As a result, since the calculated CVI value is greater than 0.591, it was accepted that the prepared rubric had a sufficient scope for measuring academic writing skills. The content validity index is the prerequisite of the construct validity process (Lawshe, 1985). It was decided that the last version of the form had a five-point rating and the use of the analytical rubric was appropriate since the performance (academic writing), which was determined, was divided into sub-dimensions (Kutlu et al., 2014).

After providing evidence for the content validity of the developed analytical rubric, the evidence for the construct validity was collected. Exploratory factor analysis (EFA) was conducted to provide evidence for construct validity. Before EFA, Kaiser-Meyer-Olkin (KMO) and Barlett sphericity tests were conducted to determine whether the relevant data set had a factorizable structure. It is stated that for a data set to be factorizable, the KMO value should be 0.70 and the Barlett sphericity test should be significant (Cokluk et al., 2012). The KMO value for the relevant data set was found to be 0.875 and the Barlett spherical test was found to be statistically significant (χ^2 (sd) = 956.427 (105); p = 0.000).

It was determined that there were no losses and misfit in the data set and the relationships between variables were linear. Test of normality was performed for each criterion and it was investigated that except two, all criteria showed normal distribution. During the process of EFA, the average score for each criterion rated by 45 raters in the experimental and control groups for the 39 student essays was analyzed. As a result of EFA analysis, it was found that the criteria were collected under a single factor and the variance was 70.05% (the factor loadings of the criteria for the relevant data set were as follows; 0.842; 0.855; 0.936; 0.968; 0.644; 0.860; 0.960; 0.987; 0.945; 0.605; 0.911; 0.891; 0.899; 0.861 and 0.622).

After collecting the evidence for the validity of the measurements obtained from the developed analytical rubric, McDonald ω coefficient was used for proof of the reliability of the measurements (McDonald, 1999). The reason for using the McDonald ω coefficient is to obtain more consistent and unbiased estimates of reliability (Osburn, 2000) in such measurements, since the factor loads of variables are different from each other (congeneric measurements). As a result of the analysis, McDonald ω coefficient was found to be 0.971 (95% Confidence Interval: 0.956-0.980). When the evidence obtained for reliability and validity is considered, it can be said that the measurements obtained from the analytical rubric to assess academic writing skills (related performance) are reliable and the inferences made based on these measurements are valid.

2.4. Experimental Procedure

In this section, information about the experimental procedure (rater training) that was applied to the experimental group was presented. First of all, all the raters in both experimental and control groups were informed about the developed analytical rubric and the performance to be assessed (academic writing skill). Moreover, they were informed about rubrics, their types, and how they were prepared the reason for this was to ensure that the experimental and control groups would have similar characteristics and experiences. In this way, it was aimed to reduce the possibility of mixing the different variance sources (variance unrelated to the structure) to the performance (related structure) to be determined. Next, the experimental and control groups were given detailed information about the analytical rubric's criteria and rating. All of these procedures took a total of three weeks, one hour each week, before the experimental procedure. In addition, the raters were not given any information as to whether they were in the experimental or control groups. After these procedures, all raters were given a 'rater file' that contained pre-prepared and numbered student essays, analytical rubric, and personal information form (pre-test), and they were given a week to assess student essays according to the developed analytical rubric. At the end of one week, rater files were collected and the assessments were transferred to the computer environment and the data set was analyzed. As a result of the analysis, it was identified that the experimental and control groups displayed similar rater behaviors in the process of assessing students' essays. Later on, the rater training was launched. Detailed information about the rater training was presented in the next section.

2.4.1. *Rater training*

To create a common structural framework (academic writing skill) among the raters in the study, rater error training (RET) and frame of reference training (FORT) were combined and applied. These two pieces of training were combined because although the RET is useful in terms of defining rater behaviors, it is not effective on rater accuracy, and the FORT is useful and effective on rater accuracy (Murphy & Balzer, 1989; Sulsky & Day, 1992). In other words, both patterns are chosen because they are complementary to each other.

The basic assumption of the RET design is that being familiar with common rater behaviors and encouraging raters to avoid these mistakes will directly lead to a decrease in rater behavior and thus more effective performance evaluation (Woehr & Huffuct, 1994). Studies have not found any evidence that RET design has a positive effect on scoring features such as inter-rater reliability (Bernardin & Pence, 1980; Borman, 1975). Although rater behaviors such as rater strictness and generosity decreased in the RET design, it was reported that scoring accuracy also decreased (Bernardin & Pence, 1980). Many researchers citing these results stated that the RET design was an inappropriate approach.

Although there are many rater training designs, it has been stated that the most preferred method is frame-of-reference traning (FORT) (Roch et al., 2012). The main reason for this is the use of a common conceptualization of performance for raters when performance is observed and evaluated (Aguinis et al., 2009; Athey & McIntyre, 1987). One reason for the effective use of the frame of reference training is that it encompasses performance theory, which is an explanation of various performance dimensions. Performance theory explains how rater behavior matches the appropriate dimension, how the effectiveness of rater behavior is evaluated, and how different judgments combine with the scoring dimension of performance (Sulsky & Day, 1992).

A rater module has been developed by the researchers for rater training. The rater training was given to student teachers of English who have taken the measurement and evaluation course for a total of four weeks and one hour each week. The rater training was implemented based on the sequence of the application in the rater module attached.

2.5. Data Analysis

Many Facet Rasch Model (MFRM) and independent samples *t*-test were used in the analysis of the data set. There are three dimensions in the study: raters, students, and criteria, and a fully crossed pattern was used because the raters assessed all students based on all the criteria.

2.5.1. *Many facet rasch model*

In the basic Rasch model, the individual and test items or performance tasks are assessed and the skill differences of the individuals and the difficulty levels of the items are placed on an equally spaced scale. It is claimed that the obtained results are independent of the sample (Sudweeks et al., 2005). In the Many Facet Rasch Model, many variability sources (such as rater, item, task, individual, time) can be placed on a single equally spaced scale (Linacre, 1993). MFRM is also known as facet models (Eckes, 2015). Although the MFRM model takes into account all variability sources, it also focuses on the interaction of these variability sources with each other (Abu Kassim, 2007). The Many Facet Rasch Model is a linear model that calibrates all parameters and converts the observations in the ranking scale to an equidistant logit scale (Bond & Fox, 2015). Logistic transformation of sequential category probabilities (log odds) enables independent variables such as peer assessment, assessment criteria, and open-ended items to be seen as dependent variables (Esfandiari, 2015). The Many Facet Rasch Model provides researchers with information that the models based on classical test theory and generalizability theory cannot provide (Lunz et al., 1990).

In this study, because academic essays written by a group of students were assessed by a group of raters, the model of the research was defined as follows:

$$\log\left(\frac{P_{bkpx}}{P_{bkpx-1}}\right) = \theta_b - \beta_k - \alpha_p - \tau_x \tag{1}$$

Pbkpx = the probability of giving an x score to a student’s certain criterion by the rater

Pbkpx−1 = the probability of giving an x-1 score to a student’s certain criterion by the rater

θb = b. the student's proficiency level,

k = k. the difficulty of the criterion,

αp = p. the severity of the rater,

τx = difficulty of getting an x score instead of x-1.

Assumptions to be met for the Many Facet Rasch Model are one-dimensionality, local independence and model data fit. First of all, when the one-dimensionality assumption is examined, it can be said that the related assumption is met, since the developed analytical rubric, as shown in the data collection tools section, has one factor. After the one-dimensionality assumption was met, G^2 statistics, which was proposed by Chen and Thissen (1997) was used to test local independence assumption. According to this statistic, the standardized LD χ^2 value estimated between each variable pair is below 10 and the marginal fit χ^2 value estimated for each variable is close to zero, which indicates local independence. In this context, estimates were made according to the generalized partial credit model and it was found that the standardized LD χ^2 values ranged from -0.4 to 4.5, and the marginal fit χ^2 values were close to zero, and as a result, local independence was achieved. Finally, standardized residual values were examined for model-data fit. For the model-data fit, it has been stated that the number of standardized residual values outside the ± 2 range should not be more than 5% of the total number of observations, and the standardized residual values outside the ± 3 range should not be more than 1% of the total data number (Linacre, 2017). Since the total number of observations for the pretest application is 39x45x15 = 26.325, the number of standardized residual values outside the ± 2 range is 1.067 (4.05%) and the number of standardized residual values outside the ± 3 range is 164 (0.62%). It was observed that model-data fit was achieved for pre-test application. The total number of observations for the post-test application was 26.322 (3 missing data), while the number of standardized residual values outside the ± 2 range was 995 (3.78%) and the number of standardized residual values outside the ± 3 range was 186 (0.71%) and it was accepted that model data fit was achieved for posttest. As a result, all the assumptions were met and the process of analysis was started.

3. FINDINGS

Findings were provided under headings as two hypotheses were tested. Besides, the measurement reports used in determining the rater errors were given in the appendices.

3.1. Research Findings of Pre-Rater Training

Literature warns that many rater errors get involved in the measurements when assessing the performance of an individual (Royal & Hecker, 2016). The present study examined the most frequently occurring rater errors such as rater severity, rater leniency, central tendency, and halo effect. Before the rater training, the rater facet measurement report given in Appendix 3 was examined for the rater severity and rater leniency errors involved in the measurements in the assessment of student compositions. Measurement reports were calculated for each facet and surface interactions (common interactions) in the MFRM. These measurement reports consisted of two parts as a group and individual levels. Measurement reports were first assessed at the group level and then at the individual level (MyFord & Wolfe, 2004). The current research considered this path.

Regarding the pre-test measurement report related to the rater facet in Appendix 2, group-level statistics (separation rate, separation index, and separation index reliability values) were found high. This indicates that the raters exhibited different errors in the process of assessing individual performance. The fixed effective chi-square value was examined to identify whether the raters showed different errors in the performance determination process, and this value was found significant ($\chi^2(44) = 2\ 835.70$; $p < .05$). After determining that different rater errors were involved in the measurements at the group level, we attempted to identify which rater or raters showed different errors in assessing the individual performance by examining the statistics at the individual level. The logit value is one of the most important statistics at the individual level. By using the logit value, the *t*-value for each rater was obtained, and this value was compared with the critical *t* value in the *t* distribution table, and the rater error was determined. As is seen in Appendix 2, the *t*-value was calculated for each rater. Since there were 22 raters in the experimental group and 23 raters in the control group, the small sample size was taken as the basis. Besides, the degree of freedom was taken as 21 and the statistical significance level was taken as α = .05. Considering *t* distribution table, the critical *t* value is 2.831. Accordingly, if the *t*-value calculated for each rater is greater than +2.831, it is assumed that the rater exhibits the leniency behavior, if it is less than -2.831, the rater exhibits the severity behavior. Figure 1 presents the graphical representation of the rater leniency and severity in the experimental and control groups that appeared in the pre-test scoring.

Figure 1. *The t-values obtained from the pre-tests of the experimental and control groups. (each point in the figure represents a rater).*

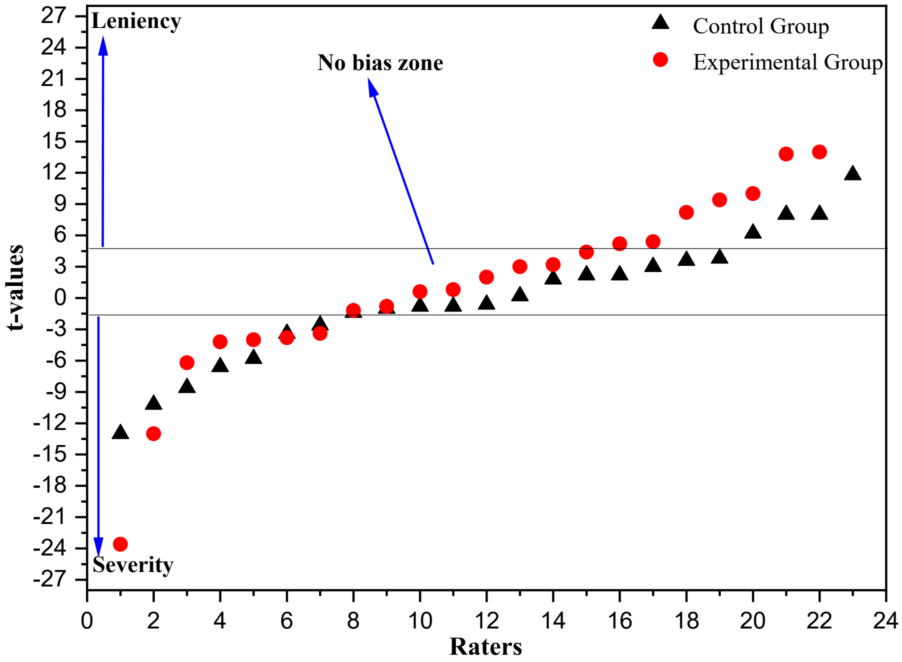


Figure 1 indicates that the raters in the control and experimental groups exhibited similar rater errors. Regarding the raters in the control group, 11 raters (47.82%) did not show leniency and severity (ideal scorer), but six raters (26.09%) showed severity, and six raters (26.09%) showed leniency. For the raters in the experimental group, five raters (22.73%) did not exhibit leniency and severity (ideal scorer), but 10 raters (45.46%) showed leniency, and seven raters (31.81%) demonstrated severity. To determine whether the logit values of the experimental and control groups differed from each other in terms of severity and leniency, independent samples t-test was conducted (see Table 2).

Table 2. *Independent samples t-test results related to the difference between pre-test scores of experimental and control groups.*

Test	Group	N	$\overline{X}$	S	*t*	*df*	*p*
Pre-test	Control	23	-0.04	0.38	0.735	43	0.466
	Experimental	22	0.05	0.43			

Note. **p* < .05 Criteria: "Control=1"; "Experimental=2"

The logit values obtained from the pre-test of the raters in the experimental and control groups were not statistically significant ($t_{(43)}$ =0.735; $p > 0.05$). In other words, both groups showed similar rater errors and were involved in the measurements at a similar rate before the rater training.

Another rater error was central tendency. To determine the central tendency involved in the measurement of the individual performance, firstly, category statistics were calculated. Table 3 presents category statistics for pre-test results.

Table 3. *Category statistics calculated for pre-test of experimental and control groups.*

Scoring categories	Frequency	%	Cumulative %	Average logit measure	Expected logit measure	Outfit
0	595	2	2	-0.14	-0.33	1.20
1	2.194	8	11	0.11	0.11	1.00
2	6.435	24	35	0.56	0.59	1.00
3	10.132	38	74	1.10	1.11	1.00
4	6.969	26	100	1.67	1.64	1.00

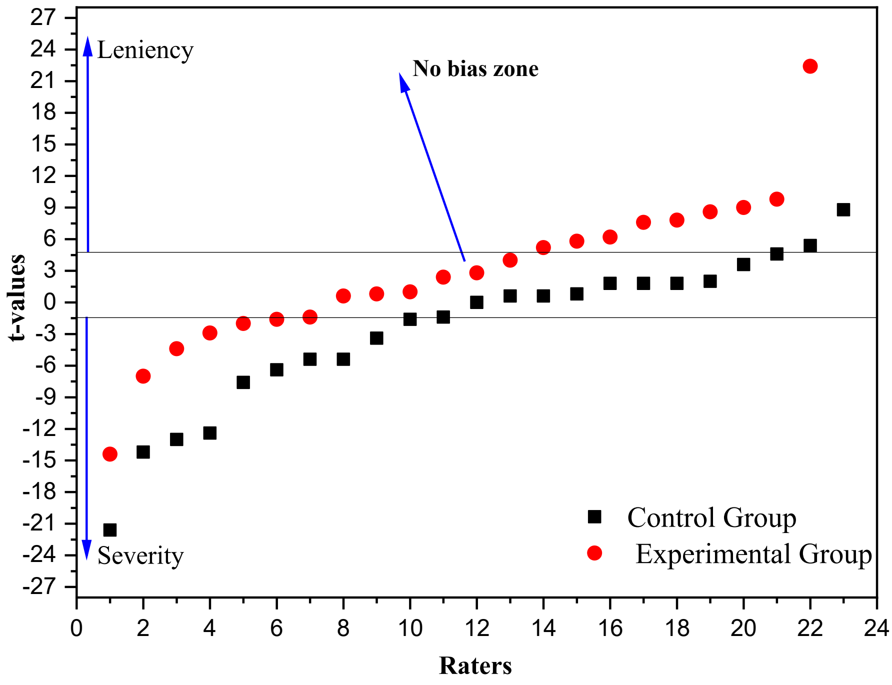
As is seen in Table 3, extreme categories were preferred less, while middle categories were preferred more. In such a case, either the raters showed central tendency or the students (whose assessment preference was determined) were at the intermediate level. Therefore, referring only to category statistics at group level does not provide enough information; other statistics should also be examined. One of these statistics is the measurement report calculated for the individual / student facet. The measurement report emphasized that separation rate, separation index and separation index reliability were high. In other words, students were successfully distinguished according to their performance levels. Besides, the significant chi-square value was interpreted as statistical evidence that students were significantly differentiated according to their performance level (χ^2 (38) = 7 695.00; p = <.05). Based on these findings, it can be said that there was no central tendency at the group level, and the current situation in category statistics was due to the performance level of the students. After determining that central tendency did not interfere with the measurements at the group level, statistics at individual level were analyzed. One of these statistics is the in-compliance and out-of-compliance values estimated for each rater. The in-compliance and out-of-compliance values given in Appendix 3 were between acceptable ranges (0.50 to 1.50). The category statistics for each rater should be examined for the final decision whether central tendency inferred with the measurements at the individual level. The category statistics were calculated for each rater, and rater 11 and 23 from the control group and rater 2, 4, 6, 9, and 14 from the experimental group were found to show central tendency during the process of determining their assessment preference at the group level.

Finally, halo effect was investigated. First, group level statistics were examined. Thus, the criterion facet measurement report was studied. The separation rate, separation index and separation index reliability were found high. This shows that the difficulty levels of the criteria were different from each other and that halo behavior did not interfere with the measurements at the group level. Accordingly, halo effect was not involved in group-level measurements. It is recommended to examine the suitability values of the raters by equalizing the criteria difficulties to determine whether halo behavior is interfered with the measurements at the individual level when assessing individual performance (Linacre, 2017). If there is a rater that fits perfectly with one or both of the fit values, it is considered to show halo behavior (İlhan, 2015; Linacre, 2017). In this context, the criterion difficulties were equalized, the analysis was repeated, and the fit values of the raters were examined. According to results, rater 2 from the control group as well as rater 17 and 22 from the experimental group showed halo effect. Linacre (2017) suggests re-examining the suitability statistics of raters by equating criterion difficulties to determine whether there is a halo effect in assessing individual performance. Therefore, MFRM analysis was repeated by setting the criterion difficulties equal to zero and the suitability statistics of the raters were examined. After the criterion difficulties were equalized, raters with a fit statistics (infit and outfit) perfectly fit with the data (with UI = 1.00 or UD = 1.00) were considered to show halo effect (Ilhan, 2015; Linacre, 2017). After the criterion difficulties were equalized, the measurement report regarding the rater aspect of the control group was obtained. Besides, in the analyses performed without equalizing the criterion difficulty, when the difference between the difficulty levels of the criteria is large, the suitability statistics are significantly greater than 1 and when the difference between the criterion difficulties is small, the relevant rater is considered to show halo behavior in performance assessment (MyFord & Wolfe, 2004). Therefore, considering the analyses performed without equalizing the criterion difficulty, it was found that three raters (1, 17 and 23) from the control group and one rater (15) from the experimental group had halo effect. As a result, it was determined that 4 raters from the control group and 3 raters from the experimental group displayed halo effect.

3.2. Research Findings of Post-Rater Training

Before rater training, upon examining rater behaviors that were involved in measurements in the process of assessing individual performance, the rater behaviors were re-examined after the rater training given. First, the effect of rater training on rater severity and rater leniency was examined. After the rater training, MFRM analysis was made for the final test, and the measurement report regarding the rater facet was presented in Appendix 3. This analysis underlined that the separation rate, separation index and separation index reliability values were high. The high values indicated that rater errors interfered with the measurements. The significance of the statistical significance test also supported this result ($\chi^2(44) = 2\ 334.60$; $p < 0.05$). After examining the statistics at the group level, the statistics at the individual level were studied. In this context, the *t*-values obtained by using the logit value were used and the *t*-value for each rater was calculated as shown in Appendix 3. Figure 2 presents the *t*-values calculated for the post-test of the experimental and control groups.

Figure 2. *t*-values obtained from the post-tests of the experimental and control groups (each point in the figure represents a rater).



As is seen in Figure 2, in the control group, 11 raters (47.83%) did not show severity or leniency, eight raters (34.78%) showed severity, and four raters (17.39%) showed leniency. In the experimental group, nine raters (40.91%) did not show severity or leniency, but 10 raters (45.46%) demonstrated leniency, and three raters (13.63%) showed severity. Besides, the raters in the experimental group got closer to the point where there was no severity and leniency. Independent samples *t*-test was conducted to determine whether the logit values obtained from the post-test of the experimental and control groups differed from each other in terms of severity and leniency. Thereby, for the effectiveness of the experimental process, the difference between raters' logit measures in the post-test and logit measures in the pre-test was taken. The differences in logit measures between post-test scores of experimental and control group was presented in Table 4.

Table 4. *Independent samples t-test result regarding the differences in logit measures between post-test scores of experimental and control groups.*

	Group	N	$\overline{X}$	*S*	*t*	*df*	*p*
Logit difference measures	Control	23	-0.09	0.16	2.708	43	0.010*
	Experimental	22	0.09	0.28			

Note. *$p < .05$ Criteria: “Control=1”; “Experimental=2”

The *t* value indicated that it was statistically significant ($t_{43} = 2.708$; $p < 0.05$; $\eta^2 = 0.15$). Based on this finding, the rater training was effective, and this effect was great. Although rater training increases the harmony between raters, it can reveal severity and leniency due to rater drift (Moore, 2009). According to the findings of the present study, after the rater training, there were drifts in the scoring of some raters; therefore, severity and leniency emerged.

Regarding the effect of rater training on central tendency, only the statistics at the individual level were examined because it was not significant at the group level before rater training. Therefore, category statistics for each rater were examined and three raters (2, 5 and 16) from the control group and one rater (number 2) from the experimental group were observed to display central tendency during the process of determining individual performance.

The effect of rater training on halo effect was examined with the statistics at the individual level. First, after the criteria difficulties were equalized, the fit statistics of each rater were examined, and one rater (17) from the control group was found to demonstrate halo effect. In the analyzes performed without equalizing the criterion difficulties, it was found that five raters (2, 5, 11, 12 and 16) from the control group and two raters (2 and 18) from the experimental group exhibited halo effect. As a result, it was found that two raters from six experimental groups from the control group displayed halo behavior in the process of assessing individual performance. In other words, six raters from the control group and two raters from the experimental group showed halo effect in the process of assessing individual performance.

4. DISCUSSION, CONCLUSION and SUGGESTIONS

Rater training was used to determine the rater errors involved in the measurements in the process of assessing individual performance and to reduce these behaviors or bring them to a controllable level. The findings were discussed under two headings in terms of before and after the experimental procedure.

4.1. Conclusions of Pre-Rater Training and Discussion

Before the rater training, it was found that raters in both the experimental and control groups displayed similar behaviors in the process of assessing individual performance. The literature emphasizes that the severity and leniency of individual performance always interfere with the measured structure during the performance assessment process (Abu Kassım, 2007; Knoch et al., 2018; Saritas-Akyol & Karakaya, 2021). Accordingly, rater's severity and leniency are important in intra-rater and inter-rater mismatches (Kane et al., 1995).

Before rater training, the raters in both groups were observed to have central tendency at the individual level (only some raters, not the whole group) while assessing individual performance. Esfandiari (2015) found that some raters showed central tendency when assessing academic writing skills, but they did not demonstrate it at the group level. A similar study was conducted by Engelhard (1994) who found that the scores of the students involved 80% of central tendency while assessing the academic writing skills. In another study, raters who did not have previous scoring experience displayed more central tendency than experienced raters (Leckie & Baird, 2011). Accordingly, the fact that the raters in both groups did not have previous scoring experience can be considered as one of the reasons for the presence of central tendency in the process of assessing the individual performance. Besides, the central tendency appeared less in performance assessment compared to severity and leniency. This indicates that the most common errors in performance assessment are severity and leniency (Cronbach, 1990).

Considering halo effect, it did not interfere with group-level measurements, but it did at the individual level. Literature advocates that halo effect is often involved in measurements and is the most studied error (Esfandiari, 2015). Engelhard (1994) also found the presence of halo effect in performance assessment. Similarly, Farrokhi and Esfandiari (2011) examined the interference of halo behavior with performance in the peer assessment, self-assessment and teacher assessment process. They observed that halo effect appeared in all three assessment types. In their study, Wu and Tan (2016) informed that some of the raters showed halo effect. In the present study, in order to prevent halo effect, the students' identity and socio-demographic information were not shared with the raters, however, halo effect was found to interfere with the measurements. This result is also supported by literature.

4.2. Conclusions of Post-Rater Training and Discussion

Before rater training, severity, leniency, central tendency and halo effect of the raters in both experimental and control groups were determined. Then, the experimental group went thorugh rater training on the aforementioned rater errors. The findings were reported based on the literature.

Considering the effect of rater training on rater severity and leniency, despite the rater training, it was found that severity and leniency were involved in the measurements during the process of assessing individual performance in both experimental and control groups. One of the reasons for this situation is thought to be the occurrence of rater drift when performance assessment spreads over time (Harik et al., 2009; Moore, 2009). The literature emphasizes that rater errors can change over time (Myford & Wolfe, 2009). When the amount of severity and leniency was examined after the rater training, it was found that the level of severity and leniency in the experimental group decreased from 77% to 59%, while severity and leniency in the control group was 52% both in the pre- and post-tests. No statistical difference was observed between the logit values showing the severity and leniency levels of the raters in the experimental and control groups before the rater training, but there was a statistical difference between the experimental and control groups after the rater training. For the practical significance of this difference, the effect size was calculated (Pallant, 2007). According to the calculated effect size value, rater training had a great effect on the rater severity and leniency, and 15% of the variability in rater severity and leniency could be explained by rater training. The literature supports this finding. Bijani (2018) found that rater training decreased the level of rater severity. Fahim and Bijani (2011) observed that rater severity and leniency involved in scoring during the assessment of students' second language writing skills decreased when rater training was given. Another study displayed that rater training had little effect on rater severity and leniency, but it had a significant effect on rater consistency (Davis, 2016). On the other hand, Weitz et al. (2014) advocated that raters scored stricter after rater training. In the study conducted by Kondo (2010), rater severity and leniency were found to be similar before and after rater training. It seems normal to have different results considering the different designs and combinations of rater training given in the literature.

When the relationship between rater training and central tendency was examined, central tendency fell from 23% to 5% in the experimental group. Therefore, it can be argued that rater training had an effect on central tendency, which is often involved in measurements when assessing individual performance. Baird et al. (2013) argued that central tendency generally occurred because of inexperienced raters who used measurement tools with multiple ratings. According to Feldman et al. (2012), if central tendency interfered with the measurements in the performance assessment process, it could jeopardize the validity of the measurements by reducing the discrimination of the individual's performance level. Accordingly, it can be interpreted that the rater training provided contributes to the validity of the measurements. May (2008) stated that rater error training design was effective in reducing central tendency. Considering the combination of rater error training and frame of reference training in the present study, the findings confirmed literature. However, Bernardin (1978) and Knoch et al. (2007) found that central tendency increased after rater training contrary to expectations.

Finally, the effect of rater training on halo behavior was examined. In the control group, while there were three raters (13.04%) with halo effect in the pre-test, this number increased to six (26.09%) in the post-test. In the experimental group, four raters (18.18%) demonstrated halo effect in the pre-test results, but it decreased to two (9.09%) in the post-test. Thus, it can be argued that rater training was effective in reducing halo effect. Feldman et al. (2012) stated that halo effect increased systematic error in performance assessment, but decreased rater accuracy, and therefore, had a significant effect on the validity of the measurements. In this context, it contributed to the validity of the measurements obtained after the rater training. Bijani (2018) found that rater training reduced halo effect. Weitz et al. (2014) stated that rater training increased raters' awareness of halo effect. In the study conducted by Pulakos (1984), rater error training design was found to be effective in reducing halo effect. Similarly, Borman (1975) concluded that rater training reduced halo effect. Accordingly, the literature supports the findings of the present study.

Based on the findings, some suggestions are as follows:

- Findings showed that one or more rater behaviors were involved in the measurements during the performance assessment processes. In this context, it is expected that the analysis and determination of rater errors in the performance assessment process will contribute to the reliability of the measurements and the validity of the inferences made from the measurements.
- Rater training was found to reduce rater errors. Accordingly, it will be beneficial to provide rater training to raters or assessors for more fair and valid measurements in the performance assessment process.
- In the present study, rater error training and frame of reference training were combined and applied. Considering that there are many rater training designs and combinations, studies can be conducted to determine more effective rater designs.
- The present study was conducted with a large group (n = 22). Literature underlines the effectiveness of smaller groups (n =5 or 6). Future studies can apply the same design in smaller groups and examine its effectiveness.

Acknowledgments

This paper was produced from the first author’s doctoral dissertation.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Gazi University, 03/04/2018, 80287700-302.08.01-54466.

Authorship Contribution Statement

Mehmet Sata: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing, original draft. **Ismail Karakaya**: Methodology, Supervision, and Validation. Authors may edit this part based on their case.

Orcid

Mehmet Sata https://orcid.org/0000-0003-2683-4997
Ismail Karakaya https://orcid.org/0000-0003-4308-6919

REFERENCES

Abu Kassim, N.L. (2011). Judging behaviour and rater errors: an application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, *11*(3), 179-197.

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.

Aguinis, H., Mazurkiewicz, M.D., & Heggestad, E.D. (2009). Using web-based frame-of-reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology*, *62*(2), 405-438. https://doi.org/10.1111/j.1744-6570.2009.01144.x

Athey, T.R., & McIntyre, R.M. (1987). Effect of rater training on rater accuracy: Levels–of–processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, *72*, 567–572. https://doi.org/10.1037/0021-9010.72.4.567

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment*, *10*(3), 1-16.

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

Bennet, J. (1998). *Human resources management*. Singapore: Prentice Hall.

Bernardin, H.J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, *63*(3), 301-308. http://dx.doi.org/10.1037 /0021-9010.63.3.301

Bernardin, H.J., & Buckley, M.R. (1981). Strategies in rater training. *Academy of Management Review,* 6(2), 205-212.

Bernardin, H.J. & Pence, E.C. (1980). Effects of rater training: New response sets and decreasing accuracy. *Journal of Applied Psychology*, *65*, 60-66. https://doi.org/10.1037/ 0021-9010.65.1.60

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education*, *5*(1), 1-20. https://doi.org/10.1080/2331186X.2018.1460901

Bond, T., & Fox, C.M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences.* Routledge. https://doi.org/10.4324/9781315814698

Borman, W.C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, *60*(5), 556-560. https://doi.org/10.1037/0021-9010.60.5.556

Brennan, R.L., Gao, X., & Colton, D.A. (1995). Generalizability analyses of work key listening and writing tests. *Educational and Psychological Measurement*, *55*(2), 157-176. https://doi.org/10.1177/0013164495055002001

Brijmohan, A. (2016). *A many-facet RASCH measurement analysis to explore rater effects and rater training in medical school admissions* [Doctoral dissertation]. https://hdl.handle.net/1807/74534

Brookhart, S.M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Brown, H.D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.

Brown, H.D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. Pearson Education.

Brown, J.D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL quarterly*, *32*(4), 653-675. https://doi.org/10.2307/3587999

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M.D. (1998). Automated scoring using a hybrid feature identification technique. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Quebec, Canada. https://doi.org/10.3115/980845.980879

Büyüköztürk, Ş. (2011). *Deneysel desenler- öntest-sontest kontrol grubu desen ve veri analizi [Experimental designs-pretest-posttest control group design and data analysis]*. Pegem Akademi.

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/10.3102/10769986022003265

Congdon, P., & McQueen, 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

Cronbach, L.I. (1990). *Essentials of psychological testing*. Harper and Row.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları* [*Multivariate statistics for social sciences: SPSS and LISREL applications*]. Pegem Akademi.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. https://doi.org/10.1177/0265532215582282

Dunbar, N.E., Brooks, C.F., & Miller, T.K. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, *31*(2), 115-128. https://doi.org/10.1007/s10755-006-9012-x

Ebel, R.L. (1965). *Measuring educational achievement*. Prentice- Hall Press.

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement*. Prentice Hall Press.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Peter Lang.

Ellis, R.O.D., Johnson, K.E., & Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly*, *36*(2), 219-233. https://doi.org/10.2307/3588333

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

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, *1*(1), 1-16.

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

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.

Feldman, M., Lazzara, E.H., Vanderbilt, A.A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, *32*(4), 279-286. https://doi.org/10.1002/chp.21156

Goodrich, H. (1997). Understanding Rubrics: The dictionary may define" rubric," but these models provide more clarity. *Educational Leadership*, *54*(4), 14-17.

Gronlund, N.E. (1977). *Constructing achievement test*. Prentice-Hall Press.

Haladyna, T.M. (1997). *Writing test items in order to evaluate higher order thinking*. Allyn & Bacon.

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

Hauenstein, N.M., & McCusker, M.E. (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment*, *25*(3), 253-266. https://doi.org/10.1111/ijsa.12177

Howitt, D., & Cramer, D. (2008). *Introduction to statistics in psychology*. Pearson Education.

Hughes, A. (2003). *Testing for language teachers.* Cambridge University Press.

İlhan, M. (2015). *Standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi [The identification of rater effects on open-ended math questions rated through standard rubrics and rubrics based on the SOLO taxonomy in reference to the many facet rasch model]* [Doctoral dissertation, Gaziantep University]. https://tez.yok.gov.tr/UlusalTezMerkezi/

İlhan, M., & Çetin, B. (2014). Rater training as a means of decreasing interfering rater effects related to performance assessment. *Journal of European Education*, *4*(2), 29-38. https://doi.org/10.18656/jee.77087

Johnson, R.L., Penny, J.A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Kane, J., Bernardin, H., Villanueva, J., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal*, *38*, 1036-1051.

Khaatri, N., Kane, M.B., & Reeve, A.L. (1995). How performance assessments affect teaching and learning. *Educational Leadership*, *53*(3), 80-83.

Kim, Y.K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model* [Doctoral dissertation, Columbia University]. https://www.proquest.com/

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.

Knoch, U., Read, J., & von Randow, T. (2007). Re-training writing raters online: How does compare with face-to-face training?, *Assessing Writing*, *12*(2), 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. *Journal of Pan-Pacific Association of Applied Linguistics*, *14*(2), 1-23.

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement*. John Wiley & Sons Incorporated.

Kutlu, Ö., Doğan, C.D., & Karaya, İ. (2014). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme [Determining student success: Determining the situation based on performance and portfolio]*. Pegem Akademi

Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Lawrence Erlbaum Associates, Inc.

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

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. (1993). Rasch-based generalizability theory. *Rasch Measurement Transaction*, *7*(1), 283-284.

Linacre, J.M. (1994). *Many-facet Rasch measurement*. Mesa Press.

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

Lumley, T., & McNamara, T.F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71. https://doi.org/10.1177/026553229501200104

Lunz, M.E., Wright, B.D. & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

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

McDonald, R.P. (1999). *Test theory: A unified approach*. Erlbaum.

McNamara, T.F. (1996). *Measuring second language performance*. Longman.

Moore, B.B. (2009). *Consideration of rater effects and rater design via signal detection theory* [Doctoral dissertation, Columbia University]. https://www.proquest.com/

Moser, K., Kemter, V., Wachsmann, K., Köver, N.Z., & Soucek, R. (2016). Evaluating rater training with double-pretest one-posttest designs: an analysis of testing effects and the moderating role of rater self-efficacy. *The International Journal of Human Resource Management*, 1-23. https://doi.org/10.1080/09585192.2016.1254102

Moskal, B.M. (2000). *Scoring rubrics: What, when and how?.*

Murphy, K.R. & Balzer, W.K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619-624. https://doi.org/10.1037/0021-9010.74.4.619

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

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.

Oosterhof, A. (2003). *Developing and using classroom assessments*. Merrill-Prentice Hall Press.

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological methods*, *5*(3), 343. http://dx.doi.org/10.1037/1082-989X.5.3.343

Pallant, J. (2007). *SPSS survival manual, a step by step guide to data analysis using spss for windows*. McGraw-Hill.

Pulakos, E.D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology*, *69*(4), 581-588. http://psycnet.apa.org/doi/10.1037/0021-9010.69.4.581

Roch, S.G., Woehr, D.J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*(2), 370-395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, *94*(1), 31-37.

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

Sarıtaş-Akyol, S., & Karakaya, İ. (2021). Investigating the consistency between students' and teachers' ratings for the assessment of problem-solving skills with many-facet Rasch measurement model. *Eurasian Journal of Educational Research*, *91*, 281-300. https://doi.org/10.14689/ejer.2021.91.13

Shale, D. (1996). Essay reliability: Form and meaning. In: White, E. Lutz, W. & Kamusikiri S. (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 76–96). MLAA.

Stamoulis, D.T. & Hauenstein, N.M.A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology*, *78*(6), 994-1003. https://doi.org/10.1037/0021-9010.78.6.994

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

Sulsky, L.M., & Day, D.V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology*, *77*(4), 501-510. https://doi.org/10.1037/0021-9010.77.4.501

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Weigle, S.C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Weitz, G., Vinzentius, C., Twesten, C., Lehnert, H., Bonnemeier, H., & König, I.R. (2014). Effects of a rater training on rating accuracy in a physical examination skills assessment. *GMS Zeitschrift für Medizinische Ausbildung*, *31*(4), 1-17.

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

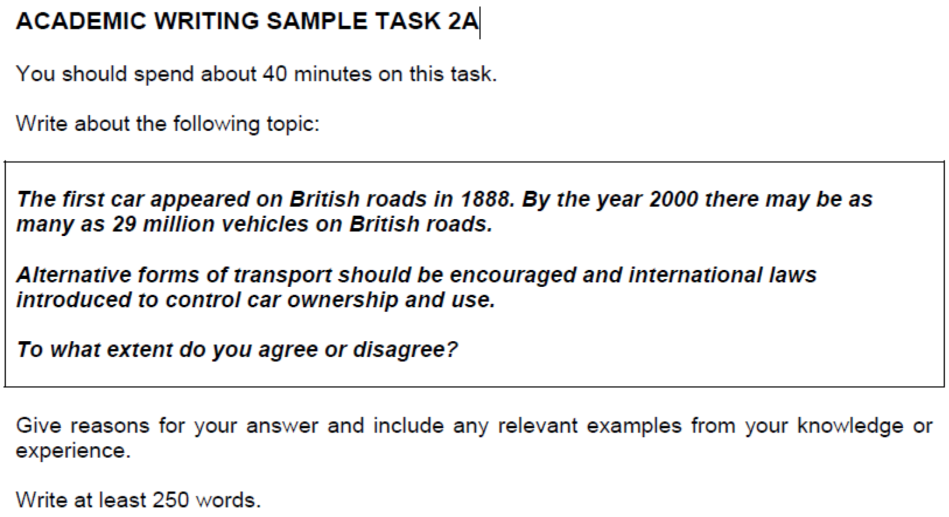
Woehr, D.J., & Huffuct, A.I. (1994). Rater training for performance appraisal. A qantitative review. *Journal of Occupational and Organizational Psychology*, *67*(3), 189-205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x

Wu, S.M., & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: the case of a university placement test. *Higher Education Research & Development*, *35*(2), 380-394. https://doi.org/10.1080/07294360.2015.1087381

Zedeck, S., & Cascio, W.F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology*, *67*(6), 752-758. https://doi.org/10.1037/0021-9010.67.6.752

APPENDIX

Appendix 1. *Academic writing sample task.*



ACADEMIC WRITING SAMPLE TASK 2A

You should spend about 40 minutes on this task.

Write about the following topic:

The first car appeared on British roads in 1888. By the year 2000 there may be as many as 29 million vehicles on British roads.

Alternative forms of transport should be encouraged and international laws introduced to control car ownership and use.

To what extent do you agree or disagree?

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

Appendix 2. *Measurement report of the rater surface of the pre-test measurements of the experimental and control groups.*

Rater Code	Logit	Standart Error	Infit	Outfit	Obs %	Exp %	Rasch Kappa	t-values
OKP01	0.09	0.05	1.26	1.26	41.60	27.60	0.193	1.80
OKP02	-0.51	0.05	1.07	1.09	40.70	27.60	0.181	-10.20
OKP03	-0.04	0.05	0.90	0.93	44.10	27.40	0.230	-0.80
OKP04	0.19	0.05	0.74	0.77	38.40	27.30	0.153	3.80
OKP05	-0.13	0.05	0.74	0.76	42.90	27.80	0.209	-2.60
OKP06	0.15	0.05	0.86	0.87	37.30	26.70	0.145	3.00
OKP07	-0.04	0.05	1.00	1.01	41.40	30.10	0.162	-0.80
OKP08	-0.03	0.05	0.79	0.82	40.80	27.80	0.180	-0.60
OKP09	-0.07	0.05	0.76	0.80	42.00	28.30	0.191	-1.40
OKP10	-0.65	0.05	1.07	1.06	33.20	28.20	0.070	-13.00
OKP11	-0.17	0.05	1.29	1.22	42.00	28.20	0.192	-3.40
OKP12	-0.05	0.05	1.29	1.29	40.50	27.70	0.177	-1.00
OKP13	-0.29	0.05	1.08	1.07	24.50	24.30	0.003	-5.80
OKP14	-0.43	0.05	0.79	0.79	42.00	28.10	0.193	-8.60
OKP15	0.11	0.05	0.84	0.85	41.40	27.80	0.188	2.20
OKP16	0.59	0.06	0.84	0.84	40.10	27.90	0.169	11.80
OKP17	0.18	0.05	1.23	1.20	44.70	28.50	0.227	3.60
OKP18	0.40	0.05	1.18	1.10	42.80	27.90	0.207	8.00
OKP19	0.01	0.05	0.76	0.86	41.50	28.20	0.185	0.20
OKP20	0.40	0.05	1.14	1.11	43.00	28.30	0.205	8.00
OKP21	0.31	0.05	0.78	0.81	43.20	28.70	0.203	6.20
OKP22	0.11	0.05	0.89	0.92	42.50	28.50	0.196	2.20
OKP23	-1.13	0.05	0.93	0.92	42.10	28.00	0.196	-6.60
ODP01	0.26	0.05	0.72	0.75	40.10	31.10	0.131	5.20
ODP02	0.41	0.05	1.36	1.43	39.10	29.30	0.139	8.20
ODP03	0.16	0.05	0.68	0.69	41.20	30.70	0.152	3.20
ODP04	0.69	0.06	1.44	1.44	44.80	30.40	0.207	13.80
ODP05	0.22	0.05	0.74	0.78	41.80	30.00	0.169	4.40
ODP06	-0.65	0.05	1.45	1.49	44.00	30.30	0.197	-13.00
ODP07	0.27	0.05	1.09	1.05	40.10	29.90	0.146	5.40
ODP08	-0.06	0.05	0.60	0.60	43.50	29.90	0.194	-1.20
ODP09	0.03	0.05	1.26	1.31	37.60	28.20	0.131	0.60
ODP10	-0.21	0.05	1.10	1.08	39.70	29.50	0.145	-4.20
ODP11	-0.31	0.05	0.73	0.87	41.30	29.30	0.170	-6.20
ODP12	-1.18	0.05	0.69	0.69	35.80	28.30	0.105	-23.60
ODP13	-0.04	0.05	0.93	0.99	40.10	27.80	0.170	-0.80
ODP14	-0.17	0.05	1.07	1.15	41.50	28.20	0.185	-3.40
ODP15	-0.20	0.05	1.49	1.47	39.70	27.50	0.168	-4.00
ODP16	0.15	0.05	0.72	0.74	41.70	27.60	0.195	3.00
ODP17	-0.19	0.05	0.91	0.92	42.80	26.70	0.220	-3.80
ODP18	0.10	0.05	1.19	1.16	39.60	27.10	0.171	2.00
ODP19	0.50	0.06	1.04	1.08	38.80	27.30	0.158	10.00
ODP20	0.04	0.05	0.86	0.93	44.50	27.60	0.233	0.80
ODP21	0.47	0.06	0.83	0.89	42.30	27.80	0.201	9.40
ODP22	0.70	0.06	1.11	1.10	23.90	24.00	-0.001	14.00
Mean	0.00	0.05	1.00	1.02				
S.D.(Population)	0.40	0.00	0.25	0.25				
S.D. (Sample)	0.40	0.00	0.25	0.25				

Model. Population : RMSE = 0.05 Adj. (True) S.D. = 0.39 Separation = 7.63
Strata = 10.51 Reliability (not inter-rater) = 0.98
Model. Sample: RMSE = 0.05 Adj. (True) S.D. = 0.40 Separation = 7.72
Strata = 10.63 Reliability (not inter-rater) = 0.98
Model. Chi-square (Fixed) : 2.835.70 d.f. = 44 *significance (probability)* = .00
Model. Chi-square (Normal) : 43.30 d.f. = 43 *significance (probability)* = .46

Note. OKP: rater who took the pre-test from the control group ODP: rater who took the pre-test from the experimental group

Appendix 3. *Measurement report of the rater surface of the pre-test measurements of the experimental and control groups.*

Rater Code	Logit	Standart Error	Infit	Outfit	Obs %	Exp %	Rasch Kappa	t-values
SKP01	0.23	0.06	1.18	1.16	46.90	30.20	0.239	4.60
SKP02	-0.71	0.05	1.33	1.34	41.80	30.90	0.158	-14.20
SKP03	0.09	0.05	0.85	0.86	45.90	31.30	0.213	1.80
SKP04	0.18	0.05	0.88	0.90	40.90	29.80	0.158	3.60
SKP05	-0.27	0.05	1.45	1.47	44.10	31.30	0.186	-5.40
SKP06	0.09	0.05	0.82	0.85	44.50	31.00	0.196	1.80
SKP07	0.03	0.05	1.27	1.19	43.30	31.80	0.169	0.60
SKP08	-0.07	0.05	0.91	0.94	45.70	31.20	0.211	-1.40
SKP09	0.03	0.05	1.08	1.11	43.70	31.00	0.184	0.60
SKP10	-0.62	0.05	1.26	1.20	44.40	31.10	0.193	-12.40
SKP11	-0.08	0.05	1.39	1.29	45.20	31.50	0.200	-1.60
SKP12	-0.27	0.05	1.45	1.41	40.90	30.90	0.145	-5.40
SKP13	-0.65	0.05	1.14	1.23	35.60	29.70	0.084	-13.00
SKP14	-0.38	0.05	0.87	0.86	46.80	31.40	0.224	-7.60
SKP15	-0.32	0.05	0.92	0.92	45.60	31.50	0.206	-6.40
SKP16	0.44	0.06	1.45	1.49	42.30	31.40	0.159	8.80
SKP17	0.04	0.05	1.05	1.07	47.20	31.70	0.227	0.80
SKP18	0.27	0.06	0.96	0.96	45.20	31.70	0.198	5.40
SKP19	-0.17	0.05	0.81	0.85	42.10	31.70	0.152	-3.40
SKP20	0.09	0.05	0.96	0.95	45.10	31.40	0.200	1.80
SKP21	0.10	0.05	1.00	0.98	43.70	31.80	0.174	2.00
SKP22	0.00	0.05	0.99	0.96	46.00	31.90	0.207	0.00
SKP23	-1.08	0.05	0.78	0.82	45.10	31.60	0.197	-21.60
SDP01	0.19	0.05	0.61	0.63	42.20	33.20	0.135	2.80
SDP02	0.26	0.06	1.40	1.49	39.70	30.20	0.136	5.20
SDP03	0.29	0.06	0.90	0.89	46.70	32.90	0.206	5.80
SDP04	1.12	0.07	1.23	1.28	44.60	32.30	0.182	22.40
SDP05	0.04	0.05	0.69	0.75	39.70	31.50	0.120	0.80
SDP06	-0.08	0.05	0.94	1.04	44.60	32.30	0.182	-1.60
SDP07	0.12	0.05	0.83	0.89	41.60	31.70	0.145	2.40
SDP08	-0.22	0.05	0.62	0.64	41.00	32.20	0.130	-4.40
SDP09	-0.16	0.05	0.61	0.62	38.90	30.40	0.122	-2.90
SDP10	0.05	0.05	1.08	1.02	42.20	31.80	0.152	1.00
SDP11	-0.35	0.05	0.87	0.96	41.50	31.00	0.152	-7.00
SDP12	-0.72	0.05	0.64	0.67	35.00	29.50	0.078	-14.40
SDP13	0.43	0.06	0.74	0.78	42.60	29.80	0.182	8.60
SDP14	-0.07	0.05	0.70	0.71	41.50	29.10	0.175	-1.40
SDP15	-0.10	0.05	1.28	1.30	39.20	29.40	0.139	-2.00
SDP16	0.20	0.05	0.70	0.73	42.80	29.30	0.191	4.00
SDP17	0.45	0.06	0.93	0.91	46.70	28.90	0.250	9.00
SDP18	0.03	0.05	1.39	1.38	38.70	28.30	0.145	0.60
SDP19	0.38	0.06	1.06	1.07	37.70	31.50	0.120	7.60
SDP20	0.39	0.06	0.84	0.84	43.60	29.40	0.201	7.80
SDP21	0.31	0.06	0.76	0.80	45.30	30.20	0.216	6.20
SDP22	0.49	0.06	1.02	1.00	24.00	23.90	0.001	9.80
Mean	0.00	0.05	1.00	1.02				
S.D.(Population)	0.38	0.00	0.24	0.24				
S.D. (Sample)	0.39	0.00	0.24	0.24				

Model. Population : RMSE = 0.05 Adj. (True) S.D. = 0.38 Separation = 7.06
Strata = 9.75 Reliability (not inter-rater) = 0.98

Model. Sample: RMSE = 0.05 Adj. (True) S.D. = 0.38 Separation = 7.14
Strata = 9.86 Reliability (not inter-rater) = 0.98

Model. Chi-Square (Fixed) : 2.334.60 d.f. = 44 *significance (probability)* = .00
Model. Chi-Square (Normal) : 43.10 d.f. = 43 *significance (probability)* = .46

Note. SKP: rater who took the post-test from the control group SDP: rater who took the post-test from the experimental group