ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 163-181

Investigating the Effect of Rater Training on Differential Rater Function in Assessing Academic Writing Skills of Higher Education Students*

Mehmet ŞATA** İsmail KARAKAYA ***

Abstract
This study aimed to examine the effect of rater training on the differential rater function (rater error) in the process of assessing the academic writing skills of higher education students. The study was conducted with a pre-test and post-test control group quasi-experimental design. The study group of the research consisted of 45 raters, of whom 22 came from experimental, and 23 came from control groups. The raters were pre-service teachers who did not participate in any rater training before, and it was investigated that they had similar experiences in assessment. The data were collected using an analytical rubric developed by the researchers and an opinion-based writing task prepared by the International English Language Testing System (IELTS). Within the scope of the research, the compositions of 39 students that were written in a foreign language (English) were assessed. Many Facet Rasch Model was used for the analysis of the data, and this analysis was conducted under the Fully Crossed Design. The findings of the study revealed that the given rater training was effective on differential rater function, and suggestions based on these results were presented.

Key Words: Academic writing, many facet Rasch model, rater training, differential rater function.

INTRODUCTION

Academic writing is defined as a type of text in which thoughts are logically structured and justified (Bayat, 2014). According to another definition, academic writing is defined as explaining the individual's views, ideas, feelings, observations, experiments, and experiences based on his/her world of thought, congruent with the rules of the language by planning them in accordance with the individual's interest towards the chosen subject (Göçer, 2010). It can be seen from these definitions that academic writing requires many skills, and it has a complex process. Academic writing consists of multiple language skills that require the use of mental, motor, and affective skills at the same time (Çekici, 2018). Essays, theses, and research reports written by students in higher education are included in academic writing types (Gillet, Hammond & Martala, 2009). Academic writing aims to convey complex thoughts, abstract concepts, and high-level mental processes (Zwiers, 2008). In this context, when academic writing is considered as the realization of higher-level mental skills, it is important to assess academic writing validly and reliably (Carter, Bishop & Kravits, 2002).

The tools that are used to assess students' academic writing skills must be authentic, which makes it difficult to choose writing tasks. Selected writing tasks need to have a place in students' lives, and if this situation is neglected, there is a risk of under-representation or a bad definition of the structure in the assessment of academic writing skills (Cumming, 2013, 2014). One of the research areas that are frequently studied in the assessment of academic writing skills is the development and assessment of students' academic writing skills in English as a second language (Aryadoust, 2016; Bitchener, Young,

* This article is derived from part of his doctoral dissertation titled "The investigation of the effect of rater training on the rater behaviors in the performance assessment process"

**Dr., Faculty of Education, Agri Ibrahim Cecen University, Turkey, mehmetwsata@gmail.com, ORCID ID: 0000-0003-2683-4997
***Prof. Dr., Faculty of Gazi Education, Gazi University, Turkey, ikarakaya2002@gmail.com, ORCID ID: 0000-0003-4308-6919

To cite this article:
Şata, M., & Karakaya, İ. (2021). *Investigating the effect of rater training on differential rater function in assessing academic writing skills of higher education students. Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 163-181. doi: 10.21031/epod.842094

Received:16.12.2020
Accepted: 30.0.2021

& Cameron, 2005; Storch & Tapper, 2009). The importance of learning a second/foreign language has been increasing every day, yet many difficulties arise in the teaching and learning process. These difficulties stem from both the complex nature of the second/foreign language learning process and the way the learning process is handled and implemented (Baştürk, 2012).

While it is important to develop students' academic writing skills, it is also important to assess these skills validly and reliably. Considering that academic writing skills are high-level mental skills, it has been stated that traditional assessment methods are not suitable; instead, performance-based assessment methods are more appropriate (Johnson, Penny & Gordon, 2008). Several features distinguish performance-based assessment from traditional assessment. While performance-based assessment has features such as being based on real-life, focusing on the process rather than the product, identifying the strong and weak skills of the individual, and prompting the individual to think more and solve problems, the traditional evaluation does not have these features (Brown & Hudson, 1998; Moore, 2009).

It can be stated that one of the important concerns about performance-based assessment is the issue of objectivity in the process of assessing individual performance and determining the situation because it is very difficult to assess objectively with performance-based assessment methods compared to traditional ones (Romagnano, 2001). Many methods have been proposed in the literature to ensure objectivity in performance-based assessment. These methods can be listed as automated scoring (Attali, Bridgeman & Trapani, 2010; Burstein et al., 1998), using more than one rater (Gronlund, 1977, p.85; Kubiszyn & Borich, 2013, p.170), using rubrics (Dunbar, Brooks & Miller, 2006; Ebel & Frisbie, 1991, p. 194; Kutlu, Doğan & Karakaya, 2014, p.51; Oosterhof, 2003, p.81), and rater training (Bernardin & Buckley, 1981; Haladyna, 1997, p.143; İlhan & Çetin, 2014; Lumley & McNamara, 1995). Each of these methods has advantages & disadvantages and strengths & weaknesses compared to each other. Haladyna (1997) emphasized that it was difficult to ensure consistency among raters, regardless of the method used. In other words, regardless of the method used, there is always the possibility that some external variables other than individual performance affect the assessments (interfere with the assessments) in performance assessment. These inconsistencies that occur in the process of assessing individual performance were defined as “rater effect/bias” (Farrokhi, Esfandiari & Vaez Dalili, 2011; Haladyna, 1997, p.139; İlhan, 2015, p.3).

In case that one or more rater errors occur during the assessment process of individual performance, the number of errors regarding the estimations of students' ability levels will be high. In other words, the estimations obtained will not be reliable. Rater errors that occur during the assessment process of individual performance also have negative effects on validity. Rater errors pose a direct validity threat since they are attributed to variance unrelated to the structure (Kassim, 2011; Brennan, Gao & Colton, 1995; Congdon & McQueen, 2000; Farrokhi et al., 2011). Therefore, it is important to minimize or control the interference of rater errors in assessments (Kim, 2009; Linacre, 1994). Rater training, which is an effective method in reducing rater errors, was used in this study (Bernardin & Buckley, 1981; Feldman, Lazzara, Vanderbilt & DiazGranados, 2012; Haladyna, 1997; Hauenstein, & McCusker, 2017; Stamoulis & Hauenstein, 1993; Weigle, 1998; Zedeck & Cascio, 1982). Rater training is widely used to reduce rater errors involved in assessments (Brijmohan, 2016). Many methods/designs regarding rater training were suggested in the literature. In this study, rater error training (RET) and frame of reference training (FRT) were used in the training of raters by combining them.

The main purpose of rater training is to enable rater to develop a common understanding of student performance and assessment criteria (Eckes, 2008; Shale, 1996). In other words, rater training ensures a valid and reliable assessment of individual performance (Moser, Kemter, Wachsmann, Köver & Soucek, 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 constant validity anxiety in the test results (Ellis, Johnson & Papajohn, 2002; McNamara, 1996). When decisions taken based on test results are vital, rater errors should be identified, and these behaviours should be reduced to an acceptable level (Ellis et al., 2002).

In statistically identifying rater errors involved in the measurements during the assessment of performance, generalizability theory and item response theory are often used. The development of

package programs in recent years has increased the frequency of using methods based on item response theory. The Rasch model, which is one of the models of item response theory, and the Many Facet Rasch Model (MFRM), which is an extension of this model, are frequently used. The main reason why MFRM is frequently used in the performance assessment process is to consider all sources of variability that are thought to affect the test scores of individuals (Kim, Park & Kang, 2012; Linacre, 1996) and to provide statistics at both individual and group level. In addition, common interactions between variability sources can be determined based on this model (Kassim, 2007). Based on these interactions, differential item functioning (DIF), differentiating individual function (DIF), and differentiating rater function (DRF) are determined (Linacre, 2017).

Differentiating rater function is defined as the tendency of the rater to give higher or lower scores to some individuals than others, depending on various characteristics of the rater, such as gender, age, and cultural factors (Wesolowski, Wind, & Engelhard, 2015). For example, a rater can give more points to successful individuals. Because the interference of differentiating rater function in the measurements is considered a systematic error, it has a negative effect on the validity of the measurements. DRF refers to a situation in which students with the same basic ability level are not likely to receive the same level of scores by raters due to their group membership. Thus, an erroneous (bias) rater prefers or dislikes a particular group of students compared to another group, for example, when scoring students' writing skills. DRF often gets involved in measurements when group memberships are known. However, in some studies, it was stated that DRF was also involved in the measurements when group membership was not known (Jin & Wang, 2017).

When the literature was examined, it was found that raters whose assessments involved severity, leniency, or central tendency error in the process of assessing individual performance, generally exhibited DRF error as well (Johnson et al., 2008; Myford & Wolfe, 2003; Wind & Guo, 2019). It was seen that studies investigating the involvement of DRF in assessing performance are quite limited. Wolfe and McVay (2012) found that 10% of the raters displayed more than one rater error in the process of assessing the essays of 120 students by 40 raters. It was investigated that some raters displayed severity, leniency, and DRF together. The study of Engelhard and Myford (2003) revealed that DRF was involved in the measurements of raters in assessing the academic writing skills of students according to their gender, race, and the language they speak. Wesolowski, Wind, and Engelhard (2015) found that DRF was involved in the measurements of 24 expert raters in assessing the jazz band performances of students. In the study conducted by Kim et al. (2012), it was found that very severe and very lenient raters generally displayed DRF. In Liu and Xie's (2014) study, 12 different scenarios were used in the process of assessing students' second language academic writing skills, and it was determined that raters showed DRF according to the scenarios. Schaefer (2008) found that errors of severity, leniency, and DRF were all involved in the process of assessing student essays. In the process of assessing performance, it was seen that rater training was used to reduce this error because DRF was frequently involved in the measurements. Bijani's (2018) study showed that the rater training given in the process of assessing students' oral presentation skills was effective. Fahim and Bijani's (2011) study revealed that rater training given in the process of assessing students' academic writing skills in the second language decreased rater x criterion interactions. On the other hand, in the study conducted by Kondo (2010), it was found that rater training given in the process of assessing second language academic writing skills did not have a significant effect on DRF. In this context, it was noticed that different results were obtained depending on the rater training pattern used and the assessed performance.

Purpose of the Study

It was observed that DRF was frequently involved in measurements in the process of assessing performances such as academic writing skills. It is significant to determine the rater errors involved in the process of assessing academic writing skills of students, such as through student essays, especially when these assessments are used in taking critical decisions such as passing a grade or getting hired in an institution. In addition, determining rater effects such as rater severity and leniency is not sufficient

by itself; it is also important to determine DRF, which is a systematic error and has a significant effect on validity. In this context, the main objective of this study is to determine the differentiating rater function and to examine the effect of rater training on DRF to provide evidence for the validity of the measurements in assessing the academic writing skills of students in higher education in second/foreign language.

METHOD

Research Design

The study was conducted with a pre-test and post-test control group quasi-experimental design (Büyüköztürk, 2011). While this pattern is an unrelated design due to the comparison of the measurements belonging to different groups, it was also defined as a relational design due to the comparison of the pre-test and post-test measurements of the same group (Howitt & Cramer, 2008).

Study Group

The research consists of a total of 45 raters, 23 from the control group and 22 from the experimental group. The raters are pre-service English teachers studying at a university's English Language Teaching Department. It was assumed that the participating pre-service teachers could assess academic writing skills since they were in the last year of their education. The average age of the raters was 21.84. A personal information form was prepared to determine whether the participants have been rater and they participated in a rater training program before, and they were asked some demographic questions. It was investigated that the participants did not participate in any rater training program before, their rating experiences were similar, and they were all inexperienced in rating. Since the efficiency of the experimental process is examined rather than the purpose of generalization to the universe in experimental studies, a universe and a sample that represents the universe have not been chosen. The scorers assessed the essays written by 39 students who were continuing their education in the first year of the same department. These students took the advanced writing and reading courses in their first year, and they were all at B1 level. The essays were collected by an academician working in the same department from the students in her course, and the students participated in the study voluntarily. While the students were writing the essays, they were informed that these essays would not be graded, and they were asked not to write their names, student numbers, or ID numbers on the papers.

Data Collection Tools

Writing task

The student essays within the scope of the research were obtained by using the opinion-based writing task published as an example by the International English Language Testing System (IELTS) (Appendix A) (IELTS, t.y.). These writing tasks are prepared in many different areas to improve students' academic writing skills in English. The main purpose here is to help students reach the level in a short time that they can write essays. These writing tasks are prepared in two different categories, academic and general, and the individual chooses one of them according to his / her area of interest. The main reason for choosing this writing task stems from the idea that it will contribute to the validity and reliability of the measurements in the process of assessing the performance of the individual since it represents real-life situations. Students were given 40 minutes for the writing task, and they were asked to write an essay consisting of at least 250 words. The essays written by the students were numbered randomly, reproduced, and distributed to the raters.Rubric (for academic writing)

In the process of assessing student essays, the analytical rubric developed by the researchers was used. A systematic process was followed in the development of the rubric, and in this way, it was aimed that it would contribute to the validity and reliability of the measurements. In this context, suggestions of

Goodrich (2000), Haladyna (1997), Kutlu et al. (2014), and Moskal (2000) were taken into consideration in the rubric development process. The literature was reviewed while determining the rubric's criteria, and sample rubrics in the studies of Weigle (2002), Hughes (2003), Brown (2004), Brown (2007), and Brookhart (2013) were comprehensively examined. After the literature review, a draft form consisting of a total of 20 sub-criteria under seven fundamental criteria was prepared, and the opinions of 11 experts in academic writing skills were consulted. The Lawshe (1975) approach was used to provide evidence for the content validity of the measurements obtained from the rubric, and the content validity rate (CVR) was calculated for each criterion. When the CVR calculated for each criterion is 0.591 and above, it was accepted that the relevant criterion has sufficient content validity (Wilson, Pan & Schumsky, 2012). In line with the opinions of the field experts, the final version of the rubric consisting of six basic criteria and 16 sub-criteria was obtained (Appendix B). Because most students did not give a title to their essays even though they were told to do it, the sub-criterion of 'Title of Essay' was not included in the many facet Rasch analysis.

After collecting the evidence for the content validity of the measures obtained from the rubric, exploratory factor analysis was performed for the construct validity. For the exploratory factor analysis, the assumptions were tested, and it was investigated that the assumptions were met (for the relevant data CVR = 0.70; χ^2 (sd) = 956.427 (105) for the Barlett sphericity test; p = 0.000). In the data set, there were no extreme values and missing data, and the relationship between the criteria was found to be linear, and except for two of them, the criteria showed a normal distribution. When the literature on how big the sample should be in the exploratory factor analysis was reviewed, it was seen that there are many different opinions. Guadagnoli and Velicer (1988) stated that all these different views were not based on a theory and that there were no experimental studies, and they emphasized that the factor loadings of the variables were important rather than the sample size in their Monte Carlo simulation study, which they conducted for the sample size required for exploratory factor analysis. Accordingly, it was stated that variables with a sample size of less than 50 people and with a factor load of 0.80 and higher, regardless of the number of variables, would produce consistent results (Guadagnoli & Velicer, 1988). Although the sample size was less than 50 participants in this study, it was found appropriate to perform an exploratory factor analysis for the data set since the factor load of all variables, except three, was greater than 0.80. Exploratory factor analysis was conducted by taking the average of the scores given by 45 raters to 39 essays. As a result of the analysis, it was found that the criteria were collected under a single factor and explained 70.05% of the variance (the factor loadings of the criteria for the relevant data set are 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).

As a result of the exploratory factor analysis, since the factor load obtained for each criterion was different (congeneric measurements), the McDonald ω coefficient (McDonald, 1999) was used for the reliability evidence of the measurements because it gave consistent results (Osburn, 2000) as a reliability determination method. As a result of the analysis, McDonald ω coefficient was found to be 0.971 (95% Confidence Interval: 0.956-0.980). Considering the reliability and validity evidence obtained for the analytical rubric, it can be argued that the measurements obtained using this measurement tool are reliable, and the inferences made based on these measurements are valid.

Experimental Process

Before starting the experimental process, to determine the starting levels of the experimental and control groups, the students' essays were distributed to the raters and the scores they gave were taken as a pre-test, and the cases of statistical differentiation were examined with the independent samples t-test and the Many Facet Rasch Model. As a result of the analysis, it was found that both groups exhibited similar rater errors in the process of assessing student essays, and the rater errors involved in the measurements were close to each other. In addition, before starting the experimental process, the analytical rubric developed for the experimental and control groups was introduced, and how to use it in the scoring process was explained. Later, both groups were explained what academic writing skill is, what its general characteristics are, and its connection with the developed rubric. These

procedures were carried out to ensure that the experimental and control groups reach a similar level at the beginning. Thus, in the process of assessing academic writing skills, the mixing of different variance sources (such as measurement tools) in the measurements was tried to be minimized. It was aimed that the raters did not know whether they were in the experimental or control group. Then, the student essays were distributed to the experimental and control groups, and they were given one week to assess the essays. One week later, student essays were collected, and they were analysed on the computer.

Rater training

To create a common understanding between raters while assessing individual performance, rater error training (RET) and frame of reference training (FRT), which are recommended in the literature, were combined. The two selected trainings were combined because of the inability of RET in defining rater behaviors and errors, but not being effective on rater accuracy, and the success of FRT on rater accuracy (Murphy & Balzer, 1989; Sulsky & Day, 1992). In other words, both rater training patterns were chosen because they were complementary to each other. The basic assumption of the RET design is that familiarity with common rater errors and encouraging raters to avoid these errors will result in a direct reduction of rater errors and, therefore, more effective performance assessment. (Woehr & Huffuct, 1994). Although rater errors such as rater severity and leniency decreased in the RET pattern, findings indicate that rating accuracy also decreases (Bernardin & Pence, 1980). In the FRT pattern, it is taken as a basis that the performance assessed is multidimensional (Selden, Sherrier & Wooters, 2012). Therefore, all sub-dimensions of performance should be defined, and behavioural examples representing these dimensions should be given to the raters. The basic principle in the FRT pattern is to train the raters to ensure that the performance dimensions assessed have certain standards. Thus, a match can be made between the scores given by the rater and the actual scores of the student (Woehr & Huffuct, 1994). The rater training was completed in four weeks in total, giving one hour each week in the measurement and evaluation course.

In the first week, the purpose, scope, and importance of rater training were introduced within the framework of RET. Then, the target audiences and the methods used were introduced in the rater training, and the first stage was completed. The second stage included information about the most common rater errors of the performance assessment process and the effects of these errors on validity and reliability. Finally, for rater training, in-group discussions were made based on a few examples. Thus, the first week of rater training was completed.

In the second week, the possible sources of rater errors involved in the measurements in the performance assessment process were explained, and the actions to be taken to reduce these errors were specified. These suggestions were determined by reviewing the literature, and the sample applications were shared with the experimental group. With this process, the RET part of the rater training was completed, and the FRT part was started. First, the academic writing skill, which was assessed by the raters, was defined. The sub-dimensions of this skill and which criteria correspond to the sub-dimensions in the rubric were explained. Then, the raters in the experimental group were asked to give representative behaviours regarding the dimensions of academic writing skill. They were then asked to discuss these representative behaviours in the group.

In the third week, as a continuation of the second week, examples regarding the dimensions of academic writing skill were given, and in-group discussions continued. After completing this stage, based on the pre-test results of the raters, the best, middle, and low-level student compositions were determined. These compositions were multiplied and distributed to the raters in the experimental group, and they were asked to be re-assessed. The raters were not informed about whether the essays were good or bad. After the assessment process, raters were randomly selected and asked about the scores they gave and the reasons for giving these scores. Later, the same question was asked to other raters in the experimental group. This process was carried out considering the criteria with the highest standard error according to the pre-test measurements. The main goal is to create a common

understanding among raters. Also, based on the pre-test measurements, written feedback was given to each rater regarding his / her ratings.

In the last week, the activities of the third week were continued with different raters. The compositions of three students, which were determined beforehand according to the pre-test results, were assessed by an academician. Raters were asked to explain how many points the field expert (academician) gave according to the determined criteria; thus, in-group discussions were made conducted. After all stages, rater training was completed, and students' compositions (39) were given to the experimental and control groups again for the post-test measurements (the duration for assessment was one week). Participation in all stages of the experimental process and scoring was voluntary. Also, additional points were added to the final grades to encourage these students.

Data Analysis

During the data analysis process, EFA and Lawshe techniques were applied in order to provide evidence for the validity of the measurements obtained from the first developed measurement tool. Then, many facet Rasch analyses were performed, and Mann Whitney U test was run based on the logit values obtained as a result of this analysis. At first, EFA was performed because the scoring of the raters showed a normal distribution. Then, since the logit values obtained by MFRM were not normally distributed, the Mann Whitney U test was used. The analysis of MFRM was preferred because it gives the common interaction between facets at the individual level. Since all raters assessed the compositions of students over all criteria, MFRM was conducted under a completely crossed-out pattern. Detailed information about MFRM was presented below.

Many facet Rasch model

MFRM has emerged as an extension of the basic Rasch model. Unlike the basic Rasch model, many variability sources (facets) such as rater, item, task, individual, time are placed on a single scale (Kim et al., 2012; Linacre, 1993; Linacre, 1996). Also, interactions between MFRM and sources of variability can be examined (Kassim, 2007). MFRM is a linear model that calibrates all parameters and converts the observations in the ranking scale to an equidistant logit scale (Bond & Fox, 2015). The logistic transformation of the log odds ratios allows independent variables such as peer assessment, status determination criteria, and open-ended items to be seen as dependent variables (Esfandiari, 2015).

Another advantage of MFRM is that it offers information that classical test theory and generalizability theory cannot provide (Lunz, Wright & Linacre, 1990). MFRM can provide the researcher with detailed information about each facet. For example, a lot of information can be obtained such as which of a group of raters assessing the performance of individuals, what the scoring is (observed value), and what the scoring should be (expected value). As MFRM provides detailed feedback, it is possible to determine which rater is good or bad and what kind of intervention is required. Based on these advantages of MFRM, the rater errors can be determined before the rater training; therefore, training can be arranged for these errors. Thus, the validity and reliability of the measurements can be increased.

Considering *rater x student composition (pxb)* interactions, the measurement model is defined as follows;

$$\ln\left(\frac{P_{bkpx}}{P_{bkpx-1}}\right) = \theta_b - \beta_k - \alpha_p - \tau_x - I_{pb} \qquad (1)$$

where

ln (P_{bkpx} / P_{bkpx-1}) = the probability that Performance b rated by Rater p on Item k in receives a rating in category x rather than category x-1,

θ_b = the logit-scale location (e.g., achievement) of Performance b,

β_k = the logit-scale location (e.g., difficulty) of Item k,

α_p = the logit-scale location (e.g., severity) of Rater p,

τ_x = the point of equal probability on the latent variable between categories

x-1 and x and

I_{pb} = Interaction term between rater facet and student composition facet.

The interaction (bias) index has an important place in determining rater errors in MFRM (Engelhard, 2002; Linacre, 2017).

Since MFRM belongs to the Rasch model family, it must meet the assumptions in the Rasch models (Eckes, 2015; Farrokhi, Esfandiari & Schaefer, 2012; Farrokhi et al., 2011). The assumptions to be met for MFRM are unidimensionality, local independence, and model data fit. As stated in the data collection tools, the rubric had a single factor structure. For the local independence assumption, the G^2 statistics proposed by Chen and Thissen (1997) were used. The standardized LD $\chi 2$ values were found to range from -0.4 to 4.5. The marginal fit χ^2 values were close to zero, and local independence was found. Standardized residual values were examined for model-data fit. The total number of observations for the pre-test application was 39x45x15 (composition x rater x criterion) = 26.325. it was observed that model-data fit was achieved for the pre-test application since the number of standardized residual values outside the ± 2 range was 1.067 (4.05%) and the number of standardized residual values outside the ± 3 range was 164 (0.62%). While the total number of observations for the post-test application was 26.322 (3 missing data), 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%).

RESULTS

Findings were presented under two headings as before (pre-test) and after (post-test) rater training. MFRM analysis was given by presenting group statistics firstly, then individual statistics.

Investigating DRF Status of Raters in Experimental and Control Groups Before Rater Training

The estimated chi-square value for the statistical indicator of *rater x student compositions (pxb)* interactions at the group level was found to be significant ($\chi 2(sd) = 5\ 298.40\ (1755)$, $p < 0.05$). According to the significance of the chi-square value, the rater function that differed at the group level was mixed up in the measurements during the assessment of student compositions. After determining that DRF was involved in the measurements at the group level in *pxb* interaction, the statistics at the individual level were examined. T statistics are used for interactions that are significant in interaction between sources of variability in MFRM. Statistical significance is tested by comparing the t-value obtained as a result of MFRM interaction analysis with the critical t-value. Interactions with a t-value outside the ± 2 range indicate differential rater function (Linacre, 2017). The number of possible interactions in the control group was 897 (23x39), and the number of significant interactions was 203 (22.63%). The number of possible interactions in the experimental group was 858 (22x39), and the number of significant interactions was 160 (18.65%). When the t statistic takes a negative value, it is defined as differential rater severity; when it takes a positive value, it refers to differential rater leniency. Table 1 presented the frequency and percentages of the raters in the experimental and control groups regarding the type of significant interactions.

Table 1. Frequencies and Percentages of Significant Interactions Regarding Pre-test Measurements in pxb Interaction

Group	Differential Rater Severity		Differential Rater Leniency		Total	
	f	%	f	%	f	%
Experimental	83	9.67	77	8.98	160	18.65
Control	111	12.37	92	10.26	203	22.63

Table 1 showed that the interference levels of the DFR of the experimental and control groups in the measurements were close to each other. The statistical significance of the differential rater severity and leniency of the raters in the control and experimental groups was tested using the bias size values obtained in MFRM interaction analysis, and analysis results were given in Table 2.

Table 2. The Results of the Mann Whitney U Test Regarding the Differentiation of Significant Interactions Regarding the Pre-test Measurements in the Experimental and Control Groups

Type of DRF	**Group**	**N**	**Average rank**	**Z**	**U**
DRS	Control	111	90.88	-1.90	3872.00
	Experimental	83	106.35		
DRL	Control	92	87.55	-0.74	3307.00
	Experimental	77	81.95		

* p<0,05; DRS = Differential Rater Severity, DRL = Differential Rater Leniency

As is seen Table 2, the interference levels of the DRF of the raters in the experimental and control groups before the rater training were statistically similar (for DRS, U = 3872.00; Z = -1.90 $p > 0.05$; for DRL, U = 3307.00; Z = -0.74; $p > 0.05$).

Investigating DRF Status of Raters in Experimental and Control Groups after Rater Training

After the experimental procedure, the estimated chi-square values for the statistical indicator of *rater x student compositions* (pxb) interactions at the group level were found to be significant ($\chi 2(sd) = 4\ 084.90\ (1755)$, $p < 0.05$). This finding shows that, despite rater training, the differential rater function in the performance assessment process of the raters interfered with the measurements.

Statistics at the individual level were examined since DRF was involved in group-level measurements. Therefore, t statistics regarding pxb interactions were examined. While 163 of 897 possible interactions (18.17%) of the control group were significant, 110 (12.82%) of 858 possible interactions of the experimental group were found to be significant. Table 3 presented the frequency and percentage values of the raters in the experimental and control groups related to the differential rater function involved in the measurements during the performance assessment process after the rater training.

Table 3. Frequency and Percentages of Significant Interactions Regarding Post-test Measurements in pxb Interaction

Group	Differential Rater Severity		Differential Rater Leniency		Toplam	
	f	%	f	%	f	%
Experimental	59	6.88	51	5.94	110	12.82
Control	95	10.59	68	7.58	163	18.17

The interference levels of the DRF of the raters in the experimental and control groups differed after the rater training while assessing student compositions. The statistical significance of the differential rater severity and leniency of the raters in the control and experimental groups was tested using the bias size values obtained in MFRM interaction analysis, and analysis results were given in Table 4.

Table 4. The Results of the Mann Whitney U Test Regarding the Differentiation of Significant Interactions Regarding the Post-test Measurements in the Experimental and Control Groups

Type of DRF	Group	N	Average rank	Z	U	*p*	d
DRS	Control	95	69.82	-2.72	2072.50*	0,007*	0.22
	Experimental	59	89.87				
DRL	Control	68	56.21	-1.38	1476.50	0,167	--
	Experimental	51	65.05				

* p<0,05; DRS = Differential Rater Severity, DRL = Differential Rater Leniency

After rater training, the interference level of the differential rater severity in the measurements in the performance assessment process was found to be statistically significant, while the interference level of the differential rater leniency was insignificant (for DRS, U = 2072.50; Z = -2.72 $p < 0.05$; for DRL, U = 1476.50; Z = -1,38; $p > 0.05$). According to this result, rater training had a small effect (r = 0.22) on differential rater severity, but no effect on differential rater leniency.

To observe the effect of rater training on pxb interactions, significant interaction numbers of the raters in the experimental group according to the pre and post-tests were given in Table 5.

Table 5**.** Significant pxb Interactions Regarding Raters in the Experimental Group

Test		P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11
Pre-test	f	9	7	7	13	7	11	7	5	13	9	2
	%	23.1	18.0	18.0	33.3	18.0	28.2	18.0	12.8	33.3	23.1	5.1
Post-test	f	2	9	10	5	4	2	2	6	6	1	8
	%	5.1	23.1	25.6	12.8	10.3	5.1	5.1	15.4	15.4	2.6	20.5
		P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22
Pre-test	f	3	9	7	13	3	5	7	5	4	6	9
	%	7.7	23.1	18.0	33.3	7.7	12.8	18.0	12.8	10.3	15.4	23.1
Post-test	f	4	5	8	8	2	2	6	8	3	3	9
	%	10.3	12.8	20.5	20.5	5.1	5.1	15.4	20.5	7.7	7.7	23.1

As is seen in Table 5, while assessing student compositions after rater training, the significant interactions of 14 raters (1, 4, 5, 6, 7, 9, 10, 13, 15, 16, 17, 18, 20, and 21) decreased (positively affected by the training); the significant interactions of 7 raters (2, 3, 8, 11, 12, 14, and 19) increased (negatively affected by the training), and the significant interactions of 1 rater (22) remained constant. To make Table 5 more understandable, the graphical representation of pxb interactions was given in Figure 1.

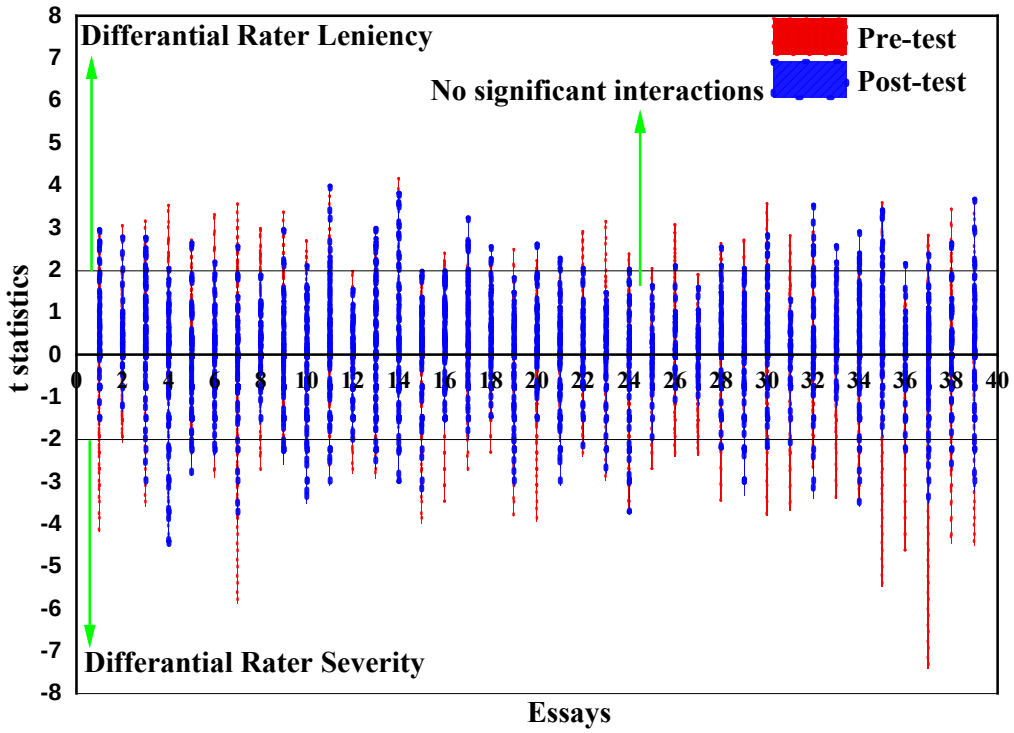


Figure 1. pxb Interactions for All Raters in the Experimental Group

As seen in Figure 1, the red lines representing the raters' pre-test were mostly outside the ± 2 range. After rater training, blue lines representing raters' ratings were observed less outside the ± 2 range. According to Figure 1, some compositions were subject to more rater bias than other compositions. For example, the raters were more severe in assessing composition numbered 37 than the other compositions. Besides, it can be said that the given rater training had a positive effect on rater errors in general, and as a result, contributed to the validity of the measurements.

DISCUSSION and CONCLUSION

This study aimed to investigate the effect of rater training on DRF, which is involved in measurements while assessing second language academic writing skills. In this context, the findings obtained before and after rater training were examined. Before rater training, DRF effect involved in the measurements was similar in both the experimental and control groups while assessing the compositions of students. Similar DRF effects were found in both group level and individual statistics. Approximately one-fifth of pxb interactions in the experimental and control groups were observed to be DRF. Research supports this finding, indicating that DRF is frequently involved in measurements in the performance assessment process (Liu & Xie, 2014; Schaefer, 2008; Wesolowski et al., 2015; Wolfe & McVay, 2012). While assessing the compositions of students, DRF involved in the measurements appeared in two ways: differential rater severity and differential rater leniency. This study found that raters mostly showed differential rater severity. The literature advocates that DRF involved in the measurements during the performance assessment process is a combination of both severity and leniency behavior, and DRF generally occurs due to too severe or too lenient raters (Kim et al., 2012). Considering that there are more severe raters in the current study, the abundance of differential rater severity confirms the literature.

During the process of assessing student compositions after the rater training, the involvement level of DRF in the measurements was examined. While the amount of change in the control group was minimal, a significant change was found in the experimental group. Although the level of interference

of the two types of DRF in the experimental and control groups in the measurements was statistically similar before the rater training, it differed statistically after the rater training. It was found that the differential rater leniency was not affected by the experimental process, but the differential rater severity was affected. In other words, rater training was effective on the differential rater severity of DRF. Considering the studies conducted by Bijani (2018), Fahim and Bijani (2011), and May (2008) and Yan (2014), rater training was effective on DRF. Van Dyke (2008) found that the differential rater leniency in the performance assessment process interfered with the measures, but the differential rater severity did not interfere. There are two main reasons for the difference between the current study and the one conducted by Van Dyke (2008): The first reason may be that the raters consisted of different groups, and the second one is that the performance assessed was different.

The results of this study can be summarized as follows;

- During the process of assessing compositions. DRF was involved in the measurements and accounted for approximately one-fifth of pxb interactions.
- Raters in the experimental and control groups exhibited similar DRF before rater training.
- Rater training had an impact on the different types of rater severity of DRF, and rater training had a small effect size on DRF.

Based on these results, some suggestions were made for future studies and researchers;

- In the present study, two different rater training patterns were combined. Considering that there are many different rater training patterns in the literature, different combinations can be made to examine the effects of rater training on DRF.
- A large experimental group was used in this study. The literature emphasizes that the training of smaller (n = 5-6) groups is more effective. Thus, it may be useful to use small groups in future studies.
- The effect of rater training on DRF can be used to train raters and contribute to the validity and reliability of the measurements during the performance assessment process utilized in placement and selection exams.

REFERENCES

Aryadoust, V. (2016). Understanding the growth of ESL paragraph writing skills and its relationships with linguistic features. *Educational Psychology*, *36*(10), 1742-1770. https://doi.org/10.1080/01443410.2014.950946

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. Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1603

Baştürk, M. (2012). İkinci dil öğrenme algılarının belirlenmesi: Balıkesir örneği. *Balikesir University Journal of Social Sciences Institute*, *15*(28-1), 251-270. Retrieved from http://dspace.balikesir.edu.tr/xmlui/handle/20.500.12462/4594

Bayat, N. (2014). Öğretmen adaylarının eleştirel düşünme düzeyleri ile akademik yazma başarıları arasındaki ilişki. *Eğitim ve Bilim*, *39*(173), 155-168. Retrieved from http://eb.ted.org.tr/index.php/EB/article/view/2333

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

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, *6*(2), 205-212. Retrieved from https://journals.aom.org/doi/abs/10.5465/amr.1981.4287782

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

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL students. *Journal of Second Language Writing*, *14*, 191–205. https://doi.org/10.1016/j.jslw.2005.08.001

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

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.* (Doktora Tezi). Retrieved from http://www.proquest.com/

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

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. New York: 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*. Ankara: Pegem Akademi Yayıncılık.

Carter, C., Bishop, J. L., & Kravits, S. L. (2002). *Keys to college studying: becoming a lifelong learner*. New Jersey: Printice Hall.

Çekici, Y. E. (2018). Türkçe'nin yabancı dil olarak öğretiminde kullanılan ders kitaplarında yazma görevleri: Yedi iklim ve İstanbul üzerine karşılaştırmalı bir inceleme. *Gaziantep Üniversitesi Eğitim Bilimleri Dergisi*, *2*(1), 1-10. Retrieved from http://dergipark.gov.tr/http-dergipark-gov-tr-journal-1517-dashboard/issue/36422/367409

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

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*(1), 1–8. https://doi.org/10.1080/15434303.2011.622016

Cumming, A. (2014). Assessing integrated skills. In A. Kunnan (Vol. Ed.), *The companion to language assessment: Vol. 1*, (pp. 216–229). Oxford, United Kingdom: Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla131

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., & Frisbie, D.A. (1991). *Essentials of educational measurement*. New Jersey: 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*. Frankfurt: 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 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*, i-60. https://doi.org/10.1002/j.2333-8504.2003.tb01893.x

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahway, NJ: Lawrence Erlbaum Associates

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. Retrieved from http://www.ijlt.ir/portal/files/401-2011-01-01.pdf

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, *34*(1), 79-101. Retrieved from https://jalt-publications.org/files/pdf-article/jj2012a-art4.pdf

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. Retrieved from https://pdfs.semanticscholar.org/dd21/ba5683dde8b616374876b0c53da376c10ca9.pdf

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

Gillet, A., Hammond, A. & Martala, M. (2009). *Successful academic writing*. New York: Pearson Longman.

Göçer, A. (2010). Türkçe öğretiminde yazma eğitimi. *Uluslararası Sosyal Araştırmalar Dergisi, 12*(3), 178-195. Retrieved from http://www.sosyalarastirmalar.com/cilt3/sayi12pdf/gocer_ali.pdf

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.* New Jersey: Prentice-Hall Press

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological bulletin*, *103*(2), 265-275. https://doi.org/10.1037/0033-2909.103.2.265

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

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*. Harlow: Pearson Education.

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

IELTS (t.y). *Prepare for IELTS*. Retrieved from https://takeielts.britishcouncil.org/prepare-test/free-sample-tests/writing-sample-test-1-academic/writing-task-2

İlhan, M. (2015). *Standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi*. (Doktora Tezi). Retrieved from https://tez.yok.gov.tr

İlhan, M., & Çetin, B. (2014). Performans değerlendirmeye karışan puanlayıcı etkilerini azaltmanın yollarından biri olarak puanlayıcı eğitimleri: Kuramsal bir analiz. *Journal of European Education*, *4*(2), 29-38. https://doi.org/10.18656/jee.77087

Jin, K. Y., & Wang, W. C. (2017). Assessment of differential rater functioning in latent classes with new mixture facets models. *Multivariate behavioral research*, *52*(3), 391-402. https://doi.org/10.1080/00273171.2017.1299615

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

Kassim, N. L. A (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. Retrieved from http://repo.uum.edu.my/3212/

Kassim, N. L. A. (2011). Judging behaviour and rater errors: an application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, 11(3), 179-197. Retrieved from http://ejournals.ukm.my/gema/article/view/49

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

Kim, Y.K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model* (Doktora Tezi). Retrieved from http://www.proquest.com/

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. Retrieved from https://eric.ed.gov/?id=EJ920513

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

Kutlu, Ö., Doğan, C.D., & Karaya, İ. (2014). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme*. Ankara: Pegem Akademi Yayıncılık.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, *28*(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J. M. (1993). Rasch-based generalizability theory. *Rasch Measurement Transaction*, *7*(1), 283-284. Retrieved from https://www.rasch.org/rmt/rmt71h.htm

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

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. *Objective measurement: Theory into practice*, *3*, 85-98. Retrieved from https://files.eric.ed.gov/fulltext/ED364573.pdf

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

Liu, J., & Xie, L. (2014). Examining rater effects in a WDCT pragmatics test. *Iranian Journal of Language Testing*, *4*(1), 50-65. Retrieved from https://cdn.ov2.com/content/ijlte_1_ov2_com/wp-content_138/uploads/2019/07/422-2014-4-1.pdf

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*. Mahwah, NJ: Erlbaum.

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

Moore, B.B. (2009). *Consideration of rater effects and rater design via signal detection theory*. (Doktora Tezi). Retrieved from http://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?*. Retrieved from http://pareonline.net/htm/v7n3.htm

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. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422. Retrieved from http://psycnet.apa.org/record/2003-09517-007

Oosterhof, A. (2003). *Developing and using classroom assessments*. New Jersey: 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

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, *94*(1), 31-37. Retrieved from http://peterliljedahl.com/wp-content/uploads/Myth-of-Objectivity2.pdf

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Selden, S., Sherrier, T., & Wooters, R. (2012). Experimental study comparing a traditional approach to performance appraisal training to a whole-brain training method at CB Fleet Laboratories. *Human Resource Development Quarterly*, *23*(1), 9-34. https://doi.org/10.1002/hrdq.21123

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). New York: 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

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes*, *8*, 207-223. https://doi.org/10.1016/j.jeap.2009.03.001

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

Van Dyke, N. (2008). Self-and peer-assessment disparities in university ranking schemes. *Higher Education in Europe*, *33*(2/3), 285-293. https://doi.org/10.1080/03797720802254114

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: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wesolowski, B. C., Wind, S. A., & Engelhard Jr, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147-170. https://doi.org/10.1177/1029864915589014

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

Wind, S. A., & Guo, W. (2019). Exploring the combined effects of rater misfit and differential rater functioning in performance assessments. *Educational and psychological measurement*, *79*(5), 962-987. https://doi.org/10.1177/0013164419834613

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

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31-37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, *31*(4), 501-527. https://doi.org/10.1177/0265532214536171

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

Zwiers, J. (2008). *Building academic language: Essential practices for content classrooms*. San Francisco: Jossey-Bass.

Appendix A. Academic Writing Sample

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 B. Rubric (For Academic Writing)

	ORGANIZATION					CONTENT	
Point	Introduction-Body-Conclusion	Thesis Statement	Topic Sentence	Supporting Sentences	Appropriate Length	Topic Relevance	Idea Development
4	The organization of introduction, body, and conclusion paragraphs is *highly* appropriate to written genre.	Thesis statement is *noticeably* given in introduction paragraph. It *comprehensively* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *comprehensively* addresses and supports the specific idea(s) given in thesis statement. It *extensively* demonstrates the main idea of the paragraph.	Supporting sentences *comprehensively* illustrate the main idea given in topic sentence.	There are *at least 250 words* in written text. It is constructed with *appropriate length*.	Written text is *highly* relevant to assigned topic in task. It *comprehensively* addresses all parts of the task.	*Extensive* details are provided to develop, support and illustrate information or ideas presented in written text.
3	The organization of introduction, body, and conclusion paragraphs is *largely* appropriate to written genre.	Thesis statement is *evidently* given in introduction paragraph. It *mostly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *mostly* addresses and supports the specific idea(s) given in thesis statement. It *largely* demonstrates the main idea of the paragraph.	Supporting sentences *adequately* illustrate the main idea given in topic sentence.	Text length is between *200 and 249 words*. It is *slightly* shorter than required length.	Written text is *mostly* relevant to assigned topic in task. It *adequately* addresses the basic parts of the task.	*Adequate* details are provided to develop, support and illustrate information or ideas presented in written text.
2	The organization of introduction, body, and conclusion paragraphs is *moderately* appropriate to written genre.	Thesis statement is *less explicitly* given in introduction paragraph. It *moderately* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *moderately* addresses and supports the specific idea(s) given in thesis statement. It demonstrates the main idea of the paragraph in *some respects.*	Supporting sentences *moderately* illustrate the main idea given in topic sentence.	Text length is between 150 *and 199 words*. It is *seemingly* shorter than required length.	Written text is *moderately* relevant to assigned topic in task. It *partially* addresses the basic parts of task.	*Basic* details are provided to develop, support and illustrate information or ideas presented in written text.
1	There is *inadequate* organization of introduction, body, and conclusion paragraphs in the written text.	Thesis statement is *vaguely* given in introduction paragraph. It *slightly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *partially* addresses and supports the specific idea(s) given in thesis statement. It *slightly* demonstrates the main idea of the paragraph.	Supporting sentences *partially* illustrate the main idea given in topic sentence.	Text length is between 100 *and 149 words*. It is *considerably* shorter than required length.	Written text is *slightly* relevant to assigned topic in task. It lacks addressing the basic parts of the task.	*Some details are* provided but they are not enough to develop, support and illustrate information or ideas presented in written text.
0	Written text lacks organization of introduction, body and conclusion paragraphs.	Thesis statement is not given in introduction paragraph or it does not include any specific idea(s) to be elaborated in the written text.	Topic sentence is not included in written text, or it does not address the thesis statement or demonstrate the main idea of the paragraph.	Written text does not include supporting sentences or they do not illustrate the main idea given in topic sentence.	Text length is *below 99 words.* It does not meet the requirement of appropriate length.	Written text is irrelevant to assigned topic in task. It fails to address the task adequately.	Information or ideas are not *thoroughly* developed, supported or illustrated in written text.

	COHERENCE	COHESION	GRAMMAR		VOCABULARY		MECHANICS	
Point	Coherence	Linking	Accuracy of Grammatical Forms	Syntactic Complexity	Word Choice	Lexical Range	Spelling	Punctuation
4	Information or ideas sequenced in paragraphs are *highly* consistent. There is a *considerably* logical progression between sentences in written text.	A *wide* range of cohesive devices used to connect ideas in written text provides a smooth transition between sentences.	All grammatical forms are *accurately* used in written text. The communication is *successfully* established.	Complex and sophisticated sentences are *extensively* used in written text in which syntactic structures are *highly* diverse.	All the words and phrases are *appropriately* used. The intended meaning is *clearly* conveyed in written text.	There is a *wide range* of vocabulary used in written text which includes *highly* sophisticated words and phrases.	All the needed spelling rules are *accurately* used in written text.	All the needed punctuation rules are *accurately* used in written text.
3	Information or ideas sequenced in paragraphs are *mostly* consistent. There is an *adequately* logical progression between sentences in written text.	An *adequate* range of cohesive devices used to connect ideas in written text provides an easy transition between sentences.	The use of the grammatical forms is *mostly accurate* in the written text. There are *few grammatical errors* which do not impede communication.	Complex and sophisticated sentences are *widely* used in written text in which syntactic structures are *adequately* diverse.	The use of words and phrases is *mostly appropriate*. There are *few* misused words or phrases which cannot obscure the intended meaning.	There is an *adequate range* of vocabulary used in written text which includes *largely* sophisticated words and phrases.	All the needed spelling rules are *mostly accurate* in written text but there are *few errors* which violate these rules.	All the needed punctuation rules are *mostly accurate* in written text but there are few errors which violate these rules.
2	Information or ideas sequenced in paragraphs are *moderately* consistent but there are some inconsistencies which *partially* interrupt logical progression between sentences.	The use of cohesive devices *at basic level* to connect ideas in written text provides a complete transition between sentences.	It is attempted to use the grammatical forms accurately in written text but there are *occasional grammatical errors* which slightly impede communication.	Complex and sophisticated sentences are *moderately* used in written text in which syntactic structures are *partially* diverse.	It is attempted to use the words and phrases appropriately but there are *occasionally* misused words or phrases which *slightly* obscure the intended meaning.	The *basic* vocabulary is used in written text which includes *moderately* sophisticated words and phrases.	It is intended to use the needed spelling rules *accurately* in written text but there are *occasional errors* which violate these rules.	It is intended to use the needed punctuation rules *accurately* in written text but there are *occasional errors* which violate these rules.
1	Paragraphs are constructed with *slightly* consistent information or ideas which interrupt logical progression and sequence between sentences.	A *limited* range of cohesive devices used to connect ideas in written text makes transition between sentences fragmentary.	The use of the grammatical forms is *generally inaccurate* in written text. There are *frequent grammatical errors* which largely impede communication.	Complex and sophisticated sentences are *slightly* used in written text in which syntactic structures are diverse to some extent.	The use of words and phrases is *generally inappropriate*. There are *frequently* misused words or phrases which *largely* obscure the intended meaning.	There is a *limited range* of vocabulary used in written text which includes *slightly* sophisticated words and phrases.	The use of the needed spelling rules is *largely* inaccurate. There are *frequent errors* which violate these rules.	The use of the needed punctuation rules is *largely* inaccurate. There are *frequent errors* which violate these rules.
0	Written text lacks consistency and logical progression between sentences.	There is an *inadequate* use of cohesive devices in written text which lacks transition between sentences.	The use of grammatical forms is *completely inaccurate* in the written text. This causes a breakdown in communication.	Written text lacks sentential complexity, sophistication and syntactic variety.	The use of vocabulary is completely inappropriate in written text. The intended message is obscured.	A repetitive vocabulary is largely used in written text which lacks sophistication.	All the needed spelling rules are *inaccurately* used in written text.	All the needed punctuation rules are *inaccurately* used in written text.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 163-181

Yükseköğretim Öğrencilerinin Akademik Yazma Becerilerinin Değerlendirilmesinde Puanlayıcı Eğitiminin Farklılaşan Puanlayıcı Fonksiyonu Üzerindeki Etkisinin İncelenmesi *

Mehmet ŞATA ** İsmail KARAKAYA ***

Öz

Bu araştırmanın amacı, yükseköğretimdeki öğrencilerin akademik yazma becerilerinin değerlendirilmesi sürecinde puanlayıcı eğitiminin farklılaşan puanlayıcı fonksiyonu (puanlayıcı hatası) üzerindeki etkisini incelemektir. Araştırma yarı deneysel desenlerden ön test son test kontrol gruplu deneysel desen ile yürütülmüştür. Araştırmanın çalışma grubu 22 deney ve 23 kontrol grubundan olmak üzere toplam 45 puanlayıcıdan oluşmaktadır. Puanlayıcılar, daha önce herhangi bir puanlayıcı eğitimine katılmayan ve puanlama deneyimlerinin benzer olduğu tespit edilen öğretmen adaylarından oluşmaktadır. Araştırma kapsamındaki veri, araştırmacılar tarafından geliştirilmiş olan analitik dereceli puanlama anahtarı ve Uluslararası İngilizce Dil Test Sistemi tarafından hazırlanmış olan düşünce temelli bir yazma görevi kullanılarak toplanmıştır. Araştırma kapsamında 39 öğrencinin ikinci dilde (İngilizce) yazmış oldukları kompozisyonlar değerlendirilmiştir. Veri analizi olarak çok yüzeyli Rasch modeli kullanılmış olup, bu analiz tamamen çaprazlanmış desen altında yürütülmüştür. Araştırmanın bulguları incelendiğinde verilen puanlayıcı eğitiminin farklılaşan puanlayıcı fonksiyonu üzerinde etkili olduğu bulunmuş ve bu sonuca dayalı öneriler sunulmuştur.

Anahtar Kelimeler: Akademik yazma, çok yüzeyli Rasch modeli, puanlayıcı eğitimi, farklılaşan puanlayıcı fonksiyonu.

GİRİŞ

Akademik yazma, düşüncelerin mantıksal olarak yapılandırılıp gerekçelendirildiği bir metin türü olarak tanımlanmaktadır (Bayat, 2014). Bir başka tanıma göre akademik yazma, bireyin düşünce temeline dayalı olarak görüş, fikir ve duygularıyla gözlem, deney ve tecrübelerini, seçilen konuyla ilgisi ölçüsünde planlayarak dilin kurallarına uygun biçimde anlatması olarak ifade edilmektedir (Göçer, 2010). Akademik yazmanın tanımlarına bakıldığında birçok beceriyi gerektirdiği ve karmaşık bir sürece sahip olduğu görülmektedir. Çünkü akademik yazma becerileri; zihinsel, devinimsel ve duyuşsal becerilerin aynı anda kullanılmasını gerektiren çoklu bir dil beceri setidir (Çekici, 2018). Yükseköğretimdeki öğrencilerin yazmış oldukları kompozisyonlar (essay), tezler ve araştırma raporları akademik yazı türlerine girmektedir (Gillet, Hammond & Martala, 2009). Akademik yazma; karmaşık düşünceleri, soyut kavramları ve üst düzey zihinsel süreçleri aktarma amacını taşımaktadır (Zwiers, 2008). Bu bağlamda akademik yazma, üst düzey zihinsel beceriyi gerçekleştirme olarak dikkate alındığında, geçerli ve güvenilir bir şekilde ölçülmesinin önemli olduğu belirtilmiştir (Carter, Bishop & Kravits, 2002).

Öğrencilerin akademik yazma becerilerinin değerlendirilmesi sürecinde kullanılan ölçme araçlarının otantik (yaşantıya dair) olması gerekmektedir. Bu durum yazma görevlerinin seçimini zorlaştırmaktadır. Bu bağlamda seçilen yazma görevlerinin öğrencilerin yaşantılarında bulunması gerekmekte bu durumun ihmal edilmesi durumunda ise akademik yazma becerilerinin ölçülmesi ve değerlendirilmesinde yapının eksik temsil edilme veya kötü bir tanım yapılma riski bulunmaktadır (Cumming, 2013, 2014). Akademik

*Bu makale "Performans değerlendirme sürecinde puanlayıcı eğitiminin puanlayıcı davranışları üzerindeki etkisinin incelenmesi" adlı doktora tezinin bir kısmından üretilmiştir.

**Dr., Eğitim Fakültesi, Ağrı İbrahim Çeçen Üniversitesi, Türkiye, mehmetwsata@gmail.com, ORCID ID: 0000-0003-2683-4997

***Prof. Dr., Gazi Eğitim Fakültesi, Gazi Üniversitesi, Türkiye, ikarakaya2002@gmail.com, ORCID ID: 0000-0003-4308-6919

Bu makaleye atıfta bulunmak için:
Şata, M., & Karakaya, İ. (2021). *Investigating the effect of rater training on differential rater function in assessing academic writing skills of higher education students. Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 163-181. doi: 10.21031/epod.842094

Geliş Tarihi: 16.12.2020
Kabul Tarihi: 30.05.2021

yazma becerilerinin değerlendirilmesinde sıklıkla çalışılan araştırma alanlarından biri de öğrencilerin ikinci dil olarak İngilizce akademik yazma becerilerinin gelişimi ve ölçülmesidir (Aryadoust, 2016; Bitchener, Young, & Cameron, 2005; Storch ve Tapper, 2009). Günümüzde giderek önemi artan ikinci dil öğrenme (yabancı dil eğitimi) konusunda birçok zorluk yaşanmaktadır. Bu zorluklar hem ikinci dil öğrenme sürecinin karmaşık özelliğinden hem de öğrenme sürecinin ele alınış ve uygulama biçiminden kaynaklanmaktadır (Baştürk, 2012).

Öğrencilerin akademik yazma becerilerinin geliştirilmesi önemli olmakla birlikte, bu becerilerin güvenilir ve geçerli bir şekilde ölçülüp değerlendirilmeleri de önemlidir. Akademik yazma becerilerinin üst düzey zihinsel beceri olduğu göz önüne alındığında, geleneksel değerlendirme yöntemlerinin uygun olmadığı, bunun yerine performans değerlendirme yöntemlerinin daha uygun olduğu belirtilmiştir (Johnson, Penny & Gordon, 2008). Performans değerlendirmeyi geleneksel değerlendirmeden ayıran birkaç özellik bulunmaktadır. Performans değerlendirme; gerçek yaşamdan kesitlere dayalı olma, üründen çok sürece odaklanma, bireyin iyi ve kötü becerilerini belirleme ve bireyi daha fazla düşünme ve problem çözmeye sevk etme gibi özelliklere sahipken, geleneksel değerlendirmede bu özellikler bulunmamaktadır (Brown & Hudson, 1998; Moore, 2009).

Performans değerlendirme ile ilgili önemli kaygılardan biri, birey performansının puanlanması ve durum belirlemesinin yapılması sürecinde objektiflik olduğu ifade edilebilir. Çünkü performans değerlendirmenin, geleneksel değerlendirme (sabit tepkili değerlendirme) gibi objektif (nesnel) bir şekilde puanlaması çok zordur (Romagnano, 2001). Alanyazında performans değerlendirmede nesnelliği sağlamak için çok sayıda yöntem önerilmiştir. Bu yöntemler; otomatik puanlama (Attali, Bridgeman & Trapani, 2010; Burstein ve diğerleri, 1998), birden fazla puanlayıcının kullanılması (Gronlund, 1977, s.85; Kubiszyn & Borich, 2013, s.170), dereceli puanlama anahtarlarının kullanılması (Dunbar, Brooks & Miller, 2006; Ebel & Frisbie, 1991, s. 194; Kutlu, Doğan & Karakaya, 2014, s.51; Oosterhof, 2003, s.81) ve puanlayıcı eğitimi (Bernardin & Buckley, 1981; Haladyna, 1997, s.143; İlhan & Çetin, 2014; Lumley & McNamara, 1995) olarak sıralanabilir. Bu yöntemlerin her birinin avantaj ve dezavantajları, birbirlerine göre üstün veya zayıf yönleri mevcuttur. Haladyna (1997) kullanılan yöntem fark etmeksizin puanlayıcılar arasında bir tutarlığın sağlanmasının zor olduğunu vurgulamaktadır. Başka bir deyişle kullanılan yöntemden bağımsız olarak performans değerlendirmede birey performansının dışındaki diğer bazı dışsal değişkenlerin her zaman için ölçümleri etkileme (ölçümlere karışma) olasılığı bulunmaktadır. Birey performansının değerlendirilme sürecinde meydana gelen bu tutarsızlıklar "puanlayıcı etkisi/yanlılığı" olarak tanımlanmaktadır (Farrokhi, Esfandiari & Vaez Dalili, 2011; Haladyna, 1997, s.139; İlhan, 2015, s.3).

Birey performansının puanlanması/değerlendirmesi sürecinde puanlayıcı hatalarından bir veya daha fazlasının ortaya çıkması durumunda, öğrencilerin yetenek düzeylerine ilişkin kestirimlerin hata miktarları yüksek olacaktır. Başka bir ifade elde edilen kestirimler güvenilir ölçümler olmayacaktır. Birey performansının puanlanması sürecinde, ortaya çıkan puanlayıcı hataları geçerlik üzerinde de olumsuz etkilere sahiptir. Puanlayıcı hataları, yapıyla ilişkisiz varyansa atfedildiğinden doğrudan bir geçerlik tehdidi oluşturmaktadır (Kassim, 2011; Brennan, Gao & Colton, 1995; Congdon & McQueen, 2000; Farrokhi ve diğerleri, 2011). Bundan dolayı puanlanma sürecindeki puanlayıcı hatalarının ölçümlere karışma düzeylerinin minimum seviyeye getirilmesi veya kontrol altına alınması önem arz etmektedir (Kim, 2009; Linacre, 1994). Bu bağlamda mevcut çalışmada puanlayıcı hatalarını azaltmada etkili bir yöntem olan puanlayıcı eğitimine başvurulmuştur (Bernardin & Buckley, 1981; Feldman, Lazzara, Vanderbilt & DiazGranados, 2012; Haladyna, 1997; Hauenstein, & McCusker, 2017; Stamoulis & Hauenstein, 1993; Weigle, 1998; Zedeck & Cascio, 1982). Puanlayıcı eğitimi, ölçümlere karışan puanlayıcı hatalarını azalmak için yaygın olarak kullanılmaktadır (Brijmohan, 2016). Alanyazında puanlayıcı eğitimi ile ilgili çok sayıda yöntem/desen önerilmektedir. Bu çalışmada puanlayıcıların eğitiminde, puanlayıcı hatası eğitimi (PHE) ve referans çerçevesi eğitimi (RÇE) desenleri birleştirilerek kullanılmıştır.

Puanlayıcı eğitiminin asıl amacı, puanlayıcıların öğrenci performansına ve durum belirleme ölçütlerine yönelik ortak bir anlayış geliştirmelerini sağlamaktır (Eckes, 2008; Shale, 1996). Başka bir deyişle, birey performansının geçerli ve güvenilir bir şekilde değerlendirilmesinin yapılmasını sağlamaktır (Moser, Kemter, Wachsmann, Köver & Soucek, 2016). Öğrencilerin açık uçlu bir sınavdan aldıkları puanlar,

hem öğrencinin performansından hem de puanlayıcının öğrenci performansını yorumlamasından oluştuğu için test sonuçlarında sabit bir geçerlik kaygısı oluşturmaktadır (Ellis, Johnson & Papajohn, 2002; McNamara, 1996). Test sonuçlarına dayalı alınan kararlar önemli olduğunda, puanlayıcı hataları belirlenmeli ve bu davranışlar kabul edilebilir bir seviyeye indirilmelidir (Ellis ve diğerleri, 2002).

Performans değerlendirme sürecinde ölçümlere karışan puanlayıcı hatalarının istatistiksel belirlemesinde sıklıkla genellebilirlik kuramı ve madde tepki kuramı kullanılmaktadır. Son yıllarda paket programların gelişmesi madde tepki kuramına dayalı yöntemlerin kullanılma sıklığını artırmıştır. Özellikle madde tepki kuramları modellerinden biri olan Rasch modeli ve bu modelin de uzantısı olan çok yüzeyli Rasch modeli (ÇYRM) sıklıkla kullanılmaktadır. ÇYRM'nin performans değerlendirme sürecinde sıklıkla kullanılmasınn temel nedeni bireylerin test puanlarını etkilediği düşünülen bütün değişkenlik kaynaklarını dikkate alınması (Kim, Park & Kang, 2012; Linacre, 1996) ve hem bireysel hem de grup düzeyinde istatistikler sağlanmasıdır. Ayrıca bu modele dayanarak değişkenlik kaynakları arasındaki ortak etkileşimler de belirlenebilmektedir (Kassim, 2007). Bu etkileşimlerden yola çıkarak değişen madde fonksiyonu (DMF), farklılaşan birey fonksiyonu (FBF) ve farklılaşan puanlayıcı fonksiyonu (FPF) tespit edilmektedir (Linacre, 2017).

Farklılaşan puanlayıcı fonksiyonu, puanlama sürecinde puanlayıcının cinsiyet, yaş, kültürel faktörler gibi çeşitli özelliklerine bağlı olarak bazı bireylere diğerlerine göre daha yüksek ya da daha düşük puanlar verme eğiliminde olması şeklinde tanımlanmaktadır(Wesolowski, Wind, & Engelhard, 2015). Örneğin bir puanlayıcı başarılı bireylere daha fazla puan verebilir. Farklılaşan puanlayıcı fonksiyonunun ölçümlere karışması sistematik hata olarak değerlendirildiğinden doğrudan ölçümlerin geçerliliği üzerinde olumsuz bir etki oluşturmaktadır. FPF, aynı temel yetenek düzeyine sahip öğrencilerin, grup üyelikleri nedeniyle puanlayıcılar tarafından aynı düzeyde derecelendirme alma olasılığının eşit olmadığı bir durumu ifade eder. Bu nedenle, hatalı (bias) bir puanlayıcı, örneğin öğrencilerin yazma becerilerini puanlarken başka bir gruba kıyasla belirli bir öğrenci grubunu tercih eder veya beğenmez. Farklılaşan puanlayıcı fonksiyonu, çoğunlukla grup üyelikleri bilindiğinde ölçümlere karışmaktadır. Fakat bazı çalışmalarda grup üyeliği bilinmediğinde de FPF'in ölçümlere karıştığı belirtilmiştir (Jin & Wang, 2017).

Alanyazın incelendiğinde, birey performansının değerlendirilmesi sürecinde değerlendirmelerinde katılık, cömertlik veya merkezi eğilim hatası karışan puanlayıcıların genellikle FPF hatasını da sergiledikleri bulunmuştur (Johnson ve diğerleri, 2008; Myford & Wolfe, 2003; Wind & Guo, 2019). Performans değerlendirme araştırmalarında FPF'in ölçümlere karışma durumunun araştırıldığı çalışmaların oldukça sınırlı olduğu görülmektedir. Wolfe ve McVay (2012) tarafından yapılan araştırmada 40 puanlayıcının 120 öğrenci kompozisyonunu değerlendirmesi sürecinde, puanlayıcıların %10'u birden fazla puanlayıcı hatasını gösterdiği bulunmuştur. Bazı puanlayıcıların katılık, cömertlik ve farklılaşan puanlayıcı fonksiyonunu birlikte gösterdikleri belirlenmiştir. Engelhard ve Myford (2003) tarafından yapılan araştırmada ise puanlayıcıların öğrencilerin cinsiyetine, ırkına ve konuştukları dile göre akademik yazma becerilerinin değerlendirilmesinde ölçümlere FPF karıştığı bulunmuştur. Wesolowski, Wind ve Engelhard (2015) tarafından yapılan araştırmada ise öğrencilerin jazz band performanslarının 24 uzman puanlayıcı tarafından değerlendirilmesinde ölçümlere FPF'in karıştığı tespit edilmiştir. Kim ve diğerleri (2012) tarafından yapılan araştırmada ise çok katı ve çok cömert puanlayıcıların genellikle FPF gösterdikleri bulunmuştur. Liu ve Xie (2014) tarafından yapılan araştırmada ise öğrencilerin ikinci dil akademik yazma becerilerinin değerlendirilmesi sürecinde 12 farklı senaryo kullanılmış ve puanlayıcıların senaryolara göre FPF gösterdikleri tespit edilmiştir. Schaefer (2008) tarafından yapılan araştırmada ise öğrenci kompozisyonlarının değerlendirilmesi sürecinde hem katılık ve cömertlik hatalarının hem de FPF'in ölçümlere karıştığı bulunmuştur. Performans değerlendirme sürecinde, FPF'in ölçümlere sıklıkla karışmasından dolayı bu hatanın azaltılmasına yönelik puanlayıcı eğitimlerine başvurulduğu görülmektedir. Bijani (2018) tarafından yapılan araştırmada öğrencilerin sözlü sunum becerilerinin değerlendirilmesi sürecinde verilen puanlayıcı eğitiminin etkili olduğu bulunmuştur. Fahim ve Bijani (2011) tarafından yapılan araştırmada ise öğrencilerin ikinci dildeki akademik yazma becerilerinin değerlendirilmesi sürecinde verilen puanlayıcı eğitiminin puanlayıcı x kriter etkileşimlerini azalttığı bulunmuştur. Diğer taraftan Kondo (2010) tarafından yapılan araştırmada ise ikinci dil akademik yazma becerilerinin değerlendirilmesi

sürecinde verilen puanlayıcı eğitiminin FPF üzerinde anlamlı bir etkiye sahip olmadığı bulunmuştur. Bu bağlamda kullanılan puanlayıcı eğitimi desenine ve ölçülen performansa bağlı olarak farklı sonuçlar elde edildiği görülmektedir.

Araştırmanın Amacı

Akademik yazma becerileri gibi performansların değerlendirilmesi sürecinde FPF'in ölçümlere sıklıkla karıştığı görülmektedir. Bu bağlamda öğrenci kompozisyonları gibi akademik yazma becerilerinin değerlendirilmesi sürecinde ve bu değerlendirmelerin sınıf geçme veya bir kuruma yerleşme gibi kritik kararların alınmasında kullanılması durumunda ölçümlere karışan puanlayıcı hatalarının belirlenmesi önem arz etmektedir. Ayrıca puanlayıcı katılığı ve cömertliği gibi puanlayıcı etkilerinin belirlenmesi tek başına yeterli olmayıp sistematik hata olan ve geçerlik üzerinde önemli bir etkisi olan FPF'in de belirlenmesi önemli görülmektedir. Bu bağlamda mevcut çalışmada yükseköğretimdeki öğrencilerin ikinci dildeki akademik yazma becerilerinin değerlendirilmesinde ölçümlerin geçerliğine kanıt sağlamak amacıyla farklılaşan puanlayıcı fonksiyonunun belirlenmesi ve puanlayıcı eğitiminin FPF üzerindeki etkisinin incelenmesi bu araştırmanın temel amacını oluşturmaktadır.

YÖNTEM

Araştırmanın Deseni

Araştırma, yarı deneysel desenlerden kontrol gruplu ön ve son test desen altında yürütülmüştür (Büyüköztürk, 2011). Bu desende farklı gruplara ait ölçümlerin karşılaştırılmasından dolayı ilişkisiz bir desen iken aynı zamanda aynı grubun ön ve son test ölçümlerinin karşılaştırılmasından dolayı da ilişkili bir desen olarak tanımlanmaktadır(Howitt & Cramer, 2008).

Çalışma Grubu

Araştırma, kontrol grubundan 23 ve deney grubundan 22 puanlayıcı olmak üzere toplam 45 puanlayıcıdan oluşmaktadır. Puanlayıcılar yükseköğretimdeki bir üniversitenin eğitim fakültesi İngiliz dili eğitimi bölümünde okuyan öğretmen adaylarından oluşmaktadır. Öğretmen adayları eğitimlerinin son döneminde olduğundan akademik yazma becerilerinin değerlendirmelerini yapabilecek uzmanlığa eriştikleri varsayılmıştır. Puanlayıcıların yaş ortalaması 21,84'dir. Puanlayıcıların puanlama deneyimleri ve daha önce herhangi bir puanlayıcı eğitimine katılıp katılmadıklarının belirlenmesi amacıyla kişisel bilgi formu hazırlanarak bazı demografik sorular sorulmuştur. Buna göre puanlayıcıların herhangi bir puanlayıcı eğitimine katılmadıkları ve puanlama tecrübelerinin benzer ve acemi düzeyde olduğu belirlenmiştir. Deneysel araştırmalarda evrene genelleme gibi bir amaçtan ziyade deneysel işlemin etkililiği incelendiğinden herhangi bir evren ve evreni temsil eden bir örneklem seçimi yoluna gidilmemiştir. Puanlayıcılar, aynı bölümüm birinci sınıfında eğitimlerine devam etmekte olan 39 öğrencinin yazmış oldukları kompozisyonları değerlendirmiştir. Bu öğrenciler ileri yazma ve okuma dersini birinci yarı yıllda almış ve hepsi B1 seviyesindedir. Öğrenci kompozisyonları ilgili bölümde görev yapan bir akademisyenin kendi dersinde öğrencilere gönüllük esasına dayalı olarak katılmaları sağlanarak toplanmıştır. Kompozisyonlar yazdırılırken herhangi bir not verme amacıyla yapılmadığını bu yüzden kişisel hiçbir bilgilerini (isim, fakülte numarası, TC numarası gibi) yazmamaları istenmiştir.

Veri Toplama Araçları

Yazma görevi

Araştırma kapsamındaki öğrenci kompozisyonları Uluslararası İngilizce Dil Test Sistemi (International English Language Testing System, IELTS) tarafından örnek olarak yayınlanan düşünce temelli deneme

görevi kullanılarak elde edilmiştir (Ek A) (IELTS, t.y.). Bu yazma görevleri öğrencilerin İngilizce akademik yazma becerilerinin geliştirilmesi amacıyla birçok farklı alanda hazırlanmaktadır. Buradaki temel amaç kısa zaman içinde öğrencinin kompozisyon yazabilme yeterliliğine ulaşmasıdır. Bu yazma görevleri akademik ve genel olmak üzere iki farklı şekilde hazırlanmakta ve birey kendi ilgi alanına göre bunlardan birini seçmektedir. Bu yazma görevinin seçilmesindeki temel neden gerçek yaşam durumunu kapsamasından dolayı birey performansının değerlendirilmesi sürecinde ölçümlerin geçerliğine ve güvenirliğine katkı sağlanacağı düşüncesinden kaynaklanmaktadır. Yazma görevi için öğrencilere 40 dakikalık süre verilmiş olup en az 250 kelimeden oluşan bir kompozisyon yazmaları istenmiştir. Öğrencilerin yazmış oldukları kompozisyonlar rasgele numaralandırılmış ve çoğaltılarak puanlayıcılara dağıtılmıştır.

Analitik dereceli puanlama anahtarı (akademik yazma için)

Öğrenci kompozisyonlarının değerlendirilmesi sürecinde araştırmacılar tarafından geliştirilmiş olan analitik dereceli puanlama anahtarı (ADPA) kullanılmıştır. APDA'nın geliştirilmesi sürecinde sistematik bir süreç izlenerek ölçümlerin geçerlik ve güvenirliklerine katkı sağlanılması amaçlanmıştır. Bu bağlamda rubrik geliştirme sürecinde Goodrich (2000), Haladyna (1997), Kutlu ve diğerleri (2014) ve Moskal'in (2000) önerileri dikkate alınmıştır. Rubriğin ölçütleri belirlenirken alanyazın taraması yapılmış olup, Weigle (2002), Hughes (2003), Brown (2004), Brown (2007) ve Brookhart (2013) tarafından yapılan araştırmalardaki örnek rubrikler kapsamlı bir şekilde incelenmiştir. Alanyazın taramasından sonra 7 temel kriter altında toplam 20 alt ölçütten oluşan taslak form hazırlanmış ve akademik yazma becerileri konusunda 11 alan uzmanın görüşlerine başvurulmuştur. Rubrikden elde edilen ölçümlerin kapsam geçerliğine kanıt sağlamak amacıyla Lawshe (1975) yaklaşımı kullanılmış ve her bir ölçüt için kapsam geçerlik oranı (KGO) hesaplanmıştır. Her bir ölçüt için hesaplanan KGO 0.591 ve üstü olduğunda ilgili ölçütün yeterli düzeyde kapsam geçerliğine sahip olduğu kabul edilmiştir (Wilson, Pan & Schumsky, 2012). Alan uzmanlarının görüşleri doğrultusunda altı temel ölçüt ve 16 alt ölçütten oluşan rubriğin son hali elde edilmiştir (Ek B). Öğrenci kompozisyonlarının büyük çoğunluğunda başlık atılmadığı için alt ölçütlerden biri olan "Kompozisyonun Başlığı (Title of Essay)" ölçütü çok yüzeyli Rasch ölçümlerinde analizlere dahil edilmemiştir.

ADPA'dan elde edilen ölçümlerin kapsam geçerliğine ilişkin kanıtlar toplandıktan sonra yapı geçerliği için açımlayıcı faktör analizi yapılmıştır. AFA için varsayımlar test edilmiş ve varsayımların sağlandığı belirlenmiştir (ilgili veri için KMO = 0.70; Barlett küresellik testi için $\chi 2$ (sd) = 956,427(105); p = 0,000). Veri setinde uç değer ve kayıp veri olmayıp ölçütler arsındaki ilişki doğrusal ve iki kriter dışındaki ölçütlerin normal dağılım gösterdiği bulunmuştur. Açımlayıcı faktör analizinde örneklem büyüklüğünün ne kadar olacağına yönelik alanyazın incelendiğinde, birçok farklı görüş bulunmaktadır. Guadagnoli ve Velicer (1988) tüm bu farklı görüşlerin bir kurama dayalı olmadığını ve deneysel çalışmalarının bulunmadığını belirtmiş ve AFA için gerekli olan örneklem büyüklünün ne olacağına yönelik yaptıkları Monte Carlo simülasyon çalışmasında örneklem sayısından ziyade değişkenlerin faktör yüklerinin önemli olduğunu vurgulamıştır. Buna göre örneklem sayısı 50 kişiden az ve değişken sayısı ne olursa olsun faktör yükü 0,80 ve daha yüksek olan değişkenlerin tutarlı sonuçlar üreteceği belirtilmiştir (Guadagnoli & Velicer, 1988). Mevcut çalışmada örneklem büyüklüğü 50 katılımcının altında olmakla birlikte değişkenlerin üçü hariç diğerlerinin faktör yükü 0,80 değerinden daha büyük olduğundan veri setine AFA yapılması uygun bulunmuştur. AFA 45 puanlayıcının 39 kompozisyona verdikleri puanların ortalaması alınarak yapılmıştır. Yapılan analiz sonucunda, ölçütlerin tek bir faktör altında toplandığı ve varyansın %70.05'ini açıkladığı bulunmuştur (ilgili veri seti için ölçütlerin faktör yükleri sırasıyla şu şekildedir; 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 ve 0,622).

AFA analizi sonucunda her bir ölçüt için elde edilen faktör yükünün farklı olmasından dolayı (konjenerik ölçümler) güvenirlik belirleme yöntemi olarak tutarlı sonuç vermesinden (Osburn, 2000) dolayı ölçümlerin güvenirlik kanıtları için McDonald ω katsayısı (McDonald, 1999) kullanılmıştır. Yapılan analizler neticesinde McDonald ω katsayısının 0,971 (%95 Güven Aralığı: 0,956-0,980) olduğu bulunmuştur. APDA için elde edilen güvenirlik ve geçerlik kanıtları dikkate alındığında bu ölçme aracı

kullanılarak elde edilen ölçümlerin güvenilir ve bu ölçümlere dayalı yapılan çıkarımların geçerli olduğu savunulabilir.

Deneysel İşlem

Deneysel işleme başlanılmadan önce deney ve kontrol gruplarının başlangıç düzeylerini belirlemek amacıyla puanlayıcılara öğrenci kompozisyonları dağıtılmış ve yaptıkları puanlamalar ön test olarak ele alınıp bağımsız örneklemler t-testi ve çok yüzeyli Rasch modeli ile istatistiksel olarak farklılaşma durumları incelenmiştir. Yapılan analizler sonucunda her iki grubun öğrenci kompozisyonlarını değerlendirme sürecinde benzer puanlayıcı hataları sergiledikleri ve ölçümlere karışan puanlayıcı hatalarının birbirine yakın oranda olduğu bulunmuştur. Ayrıca deneysel işleme başlanılmadan önce deney ve kontrol grubuna geliştirilmiş olan ADPA tanıtılmış ve puanlama sürecinde nasıl kullanılacağı açıklanmıştır. Daha sonra her iki gruba akademik yazma becerisinin ne olduğu ve genel özelliklerinin neler olduğu açıklanmış ve geliştirilen dereceli puanlama anahtarı ile bağlantısı açıklanmıştır. Bu işlemler deney ve kontrol grubunun başlangıçta benzer bir düzeye erişmesini sağlamak amacıyla yapılmıştır. Böylelikle akademik yazma becerisinin değerlendirilmesi sürecinde ölçümlere farklı varyans kaynaklarının (ölçme aracı gibi) karışması en alt düzeye indirilmeye çalışılmıştır. Puanlayıcılar deney veya kontrol grubu olduklarının bilinmemesi amaçlanmıştır. Daha sonra deney ve kontrol gruplarına çoğaltılmış olan öğrenci kompozisyonları verilmiş ve değerlendirmeleri için bir hafta süre verilmiştir. Bir hafta sonra öğrenci kompozisyonları toplanmış ve veriler bilgisayara girilerek analizler yapılmıştır.

Puanlayıcı eğitimi

Araştırmada, birey performansının değerlendirilmesi sürecinde puanlayıcılar arasında ortak bir anlayışın oluşturulması amacıyla alanyazında önerilen puanlayıcı eğitimlerinden puanlayıcı hatası eğitimi (PHE) ve referans çerçevesi eğitimi (RÇE) birleştirilerek uygulanmıştır. Seçilen iki eğitimin birleştirilmesinde, PHE'nin puanlayıcı davranışlarını hatalarını tanımlamasındaki başarısına rağmen puanlayıcı doğruluğu üzerinde etkili olamaması ve RÇE'in ise puanlayıcı doğruluğu üzerindeki başarısından kaynaklanmaktadır (Murphy & Balzer, 1989; Sulsky & Day, 1992). Başka bir deyişle her iki puanlayıcı eğitimi deseni birbirinin tamamlayıcısı olduğundan dolayı seçilmiştir. PHE desenin temel varsayımı, yaygın puanlayıcı hatalarına aşina olmanın ve bu hatalardan kaçınmak için puanlayıcıları teşvik etmenin puanlayıcı hatalarının doğrudan azalmasına ve dolaysıyla daha etkili performans değerlendirmeye neden olacağı şeklindedir (Woehr & Huffuct, 1994). PHE deseninde, puanlayıcı katılığı ve cömertliği gibi puanlayıcı hataları azalmasına rağmen, puanlama doğruluğunun da azaldığına yönelik bulgular rapor edilmiştir (Bernardin & Pence, 1980). RÇE deseninde ise değerlendirilmesi yapılan performansın çok boyutlu olduğu temel alınmaktadır (Selden, Sherrier & Wooters, 2012). Bu nedenle, performansın tüm alt boyutları tanımlanmalı ve bu boyutların temsilcisi olan davranışsal örnekler puanlayıcılara verilmelidir. RÇE deseninde temel ilke, puanlayıcıların değerlendirilmesi yapılan performans boyutlarının belli standartlara sahip olduğuna yönelik eğitilmesidir. Böylece puanlayıcının verdiği puanlar ile öğrencinin gerçek puanları arasında bir eşleştirme yapılabilecektir (Woehr & Huffuct, 1994). Puanlayıcı eğitimi, ölçme ve değerlendirme dersinde ve her hafta bir saat verilerek toplam dört haftada tamamlanmıştır.

İlk hafta, PHE eğitiminin çerçevesinde puanlayıcı eğitiminin amaç ve kapsamı ve önemi hakkında bilgi verilmiştir. Daha sonra puanlayıcı eğitiminde hedef kitlelerin kimler olduğu, hangi yöntemlerin kullanıldığı belirtilmiş ve ilk aşama tamamlanmıştır. İkinci aşamada ise performans değerlendirme sürecinde ölçümlere en fazla karışan puanlayıcı hatalarının neler olduğu ve bu hataların geçerlik ve güvenirlik üzerindeki etkilerinin neler olduğu tanımlanmıştır. Son olarak birkaç örnek üzerinden puanlayıcı hataları için grup içi tartışmalar yapılmış ve puanlayıcı eğitiminin birinci haftası tamamlanmıştır.

İkinci haftada ise performans değerlendirme sürecinde ölçümlere karışan puanlayıcı hatalarının olası kaynakları anlatılmış ve bu hataları azaltmak için yapılması gereken işlemler belirtilmiştir. Bu öneriler

alanyazın taraması yapılarak belirlenmiş ve örnek uygulamalar deney grubuyla paylaşılmıştır. Bu işlem ile birlikte puanlayıcı eğitimin PHE kısmı tamamlanmış ve RÇE kısmına geçilmiştir. İlk olarak puanlayıcılar tarafından değerlendirilmesi yapılan akademik yazma becerisi tanımlanmıştır. Bu beceri tanımlaması yapılırken hangi alt boyutlara sahip olduğu belirtilmiş ve alt boyutlarının hazırlanan dereceli puanlama anahtarındaki hangi ölçüte karşılık geldiği belirtilmiştir. Daha sonra deney grubundaki puanlayıcılardan akademik yazma becerisinin boyutlarına ilişkin temsil edici davranışlar vermeleri istenmiştir. Daha sonra bu temsil edici davranışlara yönelik grup içi tartışma yapmaları istenmiştir.

Üçüncü haftada ise ikinci haftanın devamı niteliğinde akademik yazma becerisinin boyutlarına ilişkin örnekler verilerek grup içi tartışmalara devam edilmiştir. Bu aşama tamamlandıktan sonra puanlayıcıların öntest sonuçlarından yola çıkarak en iyi orta ve düşük düzeydeki üç öğrenci kompozisyonları belirlenmiş ve bu kompozisyonlar çoğaltılarak deney grubundaki puanlayıcılara dağıtılmış ve yeniden değerlendirilmeleri istenmiştir. Puanlayıcılara öğrenci kompozisyonların iyi veya kötü olması hakkında herhangi bir şey söylenmemiştir. Puanlayıcıların değerlendirmeleri bittikten sonra rasgele puanlayıcılar seçilip verdikleri puanlar ve bu puanları verme nedenleri sorulmuştur. Daha sonra ise aynı soru deney grubundaki diğer puanlayıcılara da sorulmuştur. Bu işlem tüm ölçütler üzerinden olmayıp ön test ölçümlerine göre standart hatası en yüksek olan ölçütler üzerinden yapılmıştır. Burada temel amaç puanlayıcılar arasında ortak bir anlayış oluşturmaktır. Ayrıca ön test ölçümlerinden yola çıkarak her bir puanlayıcıya yaptığı puanlamalara ilişkin geri dönütler de yazılı olarak verilmiştir.

Son haftada ise üçüncü haftada yapılan etkinlikler, farklı puanlayıcılar seçilerek devam edilmiştir. Ayrıca ön test sonuçlarına göre önceden belirlenmiş olan üç öğrenci kompozisyonu alan uzmanı bir akademisyen tarafından değerlendirilmiş ve puanlayıcılara alan uzmanının belirlenen kriterlere göre kaç puan verdiği sorulmuş ve grup içi tartışmalar yapılması sağlanmıştır. Tüm aşamalardan sonra puanlayıcı eğitimi tamamlanmış ve son test ölçümleri için deney ve kontrol grubuna yeniden öğrenci kompozisyonları (39 tane) verilmiş ve değerlendirmeleri için bir hafta süre verilmiştir. Deneysel sürecin tüm aşamalarına ve puanlamaya katılım gönüllü olmuş ve bu öğrencilerin teşvik edilmesi için final notlarına ek puanlar eklenmiştir.

Verilerin Analizi

Verilerin analiz sürecinde ilk olarak geliştirilen ölçme aracından elde edilen ölçümlerin geçerliğine kanıt sağlamak amacıyla AFA ve Lawshe tekniği uygulanmıştır. Yapılan analizler sonrasında çok yüzeyli Rasch analizleri gerçekleştirilmiş ve bu analiz sonucunda elde edilen logit değerleri üzerinden fark testlerinden biri olan Mann Whitney U testi yapılmıştır. Başlangıçta puanlayıcıların yaptıkları puanlamaların dağılımı normal dağılım gösterdiğinden AFA yapılmış, daha sonra ÇYRM modeli ile elde edilen logit değerleri normal dağılmadığı için Mann Whitney U testi kullanılmıştır. Araştırmada bireysel düzeyde yüzeyler arasındaki ortak etkileşimi verdiğinden dolayı ÇYRM analizi tercih edilmiştir. Araştırma kapsamındaki tüm puanlayıcılar öğrenci kompozisyonlarını tüm ölçütler üzerinden değerlendirdiğinden ÇYRM tamamen çaprazlanmış desen altında yürütülmüştür. ÇYRM'a ilişkin detaylı bilgiler aşağıda sunulmuştur.

Çok yüzeyli rasch modeli (ÇYRM)

Çok yüzeyli Rasch modeli temel Rasch modelinin uzantısı olarak ortaya çıkmıştır. ÇYRM'de temel Rasch modelinin aksine puanlayıcı, madde, görev, birey, zaman gibi birçok değişkenlik kaynağı (yüzey-facet) tek bir ölçek üzerine yerleştirilmektedir (Kim ve diğerleri, 2012; Linacre, 1993; Linacre, 1996). Aynı zamanda MFRM ile değişkenlik kaynakları arasındaki etkileşimler de incelenebilmektedir (Kassim, 2007). Çok yüzeyli Rasch ölçme modeli bütün parametreleri kalibre ederek sıralama ölçeğindeki gözlemleri eşit aralıklı bir logit ölçeğine dönüştüren doğrusal bir modeldir (Bond & Fox, 2015). Ardışık kategori olasılıkları (log odds) oranlarının lojistik dönüşümü; akran değerlendirme, durum belirleme kriterleri ve açık uçlu madde gibi bağımsız değişkenlerin, bağımlı değişken olarak görülmesine imkân sağlamaktadır (Esfandiari, 2015).

Klasik test kuramı ve genellenebilirlik kuramının sağlayamadığı bilgileri sunması ÇYRM'nin başka bir avantajıdır (Lunz, Wright & Linacre, 1990). ÇYRM araştırmacıya her bir yüzey ile ilgili detaylı bilgi verebilmektedir. Örneğin bireylerin performansını değerlendiren bir grup puanlayıcıdan hangisinin yaptığı puanlamanın ne olduğu (gözlenen değer) ve yaptığı puanlamanın ne olması (beklenen değer) gerektiği gibi birçok bilgi elde edilebilmektedir. ÇYRM modelinin detaylı geri dönüt sağlaması sonucu hangi puanlayıcının iyi hangisinin kötü olduğu ve buna bağlı olarak nasıl bir müdahale gerektiği belirlenebilmektedir. ÇYRM'nin bu avantajlarından yola çıkarak puanlayıcı eğitimi öncesinde puanlayıcı hatalarının belirlenmesi ve bu hatalara yönelik eğitimler vermesi sağlanabilir. Böylelikle ölçümlerin geçerliği ve güvenirliği artırılabilir.

Araştırmada *puanayıcı x öğrenci kompozisyonu* (*pxb*) etkileşimleri incelendiğinden ölçme modeli aşağıdaki gibi tanımlanmaktadır;

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

Burada; P_{bkpx}, p. puanlayıcı tarafından b. öğrencinin k. kriterine x puanının verilme olasılığı, P_{bkpx-1} , p. puanlayıcı tarafından b. öğrencinin k. kriterine x-1 puanının verilme olasılığı, θ_b, b. öğrencinin yeterlilik düzeyi, β_k, k. kriterin zorluk derecesi, α_p = p. puanlayıcının katılık derecesi, τ_x, x-1 puanı yerine x puanının alınma zorluğu ve I_{pb} ise puanlayıcı yüzeyi ile öğrenci kompozisyonu yüzeyi arasındaki etkileşim terimi olarak adlandırılmaktadır. Çok yüzeyli Rasch modelinde puanlayıcı hatalarının belirlenmesinde etkileşim (yanlılık) indeksi önemli bir yer tutmaktadır (Engelhard, 2002; Linacre, 2017).

ÇYRM, Rasch modelleri ailesine ait olduğundan, Rasch modellerindeki varsayımları karşılaması gerekmektedir (Eckes, 2015; Farrokhi, Esfandiari & Schaefer, 2012; Farrokhi vd., 2011). Çok yüzeyli Rasch ölçme modeli için karşılanması gereken varsayımlar; tek boyutluluk, yerel bağımsızlık ve model veri uyumudur. Veri toplama araçlarında belirtildiği gibi dereceli puanlama anahtarı tek faktörlü bir yapıya sahiptir. Yerel bağımsızlık varsayımı için Chen ve Thissen (1997) tarafından önerilen G^2 istatistiği kullanılmış ve standartlaştırılmış LD χ^2 değerlerinin -0.4 ile 4.5 aralığında değiştiği ve marjinal uyum χ^2 değerlerinin ise sıfıra yakın olduğu bulunmuş ve yerel bağımsızlığın sağlandığı tespit edilmiştir. Model veri uyumu için standartlaştırılmış artık değerler incelenmiştir. Ön test uygulaması için toplam gözlem sayısı 39x45x15 (kompozisyonxpuanlayıcıxölçüt) = 26.325 iken ±2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı 1.067 (%4,05) ve ±3 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı ise 164 (%0,62) olduğundan ön test uygulaması için model-veri uyumunun sağlandığı görülmüştür. Son test uygulaması için toplam gözlem sayısı 26.322 (3 kayıp veri) iken ±2 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı 995 (%3,78) ve ±3 aralığının dışında kalan standartlaştırılmış artık değerlerin sayısı ise 186 (%0,71) olduğu belirlenmiştir.

BULGULAR

Araştırma kapsamında elde edilen bulgular puanlayıcı eğitimi öncesi (öntest) ve sonrası (sontest) olarak iki başlık halinde sunulmuştur. ÇYRM analiz sonuçları öncelikle grup daha sonra bireysel istatistikler sunularak verilmiştir.

Puanlayıcı Eğitimi Öncesi Deney ve Kontrol Grubundaki Puanlayıcıların FPF Durumlarının İncelenmesi

Puanlayıcı x öğrenci kompozisyonları (pxb) etkileşimlerinin grup düzeyindeki istatistiksel göstergesine ilişkin kestirilen ki-kare değerinin anlamlı olduğu bulunmuştur ($\chi2(sd) = 5\ 298.40\ (1755)$, $p < 0.05$). Ki-kare değerinin anlamlı çıkması öğrenci kompozisyonlarının değerlendirilmesi sürecinde ölçümlere grup düzeyinde farklılaşan puanlayıcı fonksiyonun karıştığına işaret etmektedir. *pxb* etkileşiminde grup düzeyinde FPF'in ölçümlere karıştığı belirlendikten sonra bireysel düzeydeki istatistikler incelenmiştir. ÇYRM'de değişkenlik kaynakları arasındaki etkileşimde anlamlı çıkan etkileşimler için t istatistiği kullanılmaktadır. ÇYRM etkileşim analizi sonucunda elde edilen t-değeri kritik t-değeri ile

karşılaştırılarak istatistiksel anlamlılık test edilmektedir. t değeri ±2 aralığının dışında olan etkileşimlerin farklılaşan puanlayıcı fonksiyonuna işaret ettiği belirtilir (Linacre, 2017). Kontrol grubunda olası etkileşim sayısı 897 (23x39) ve anlamlı bulunan etkileşim sayısı ise 203 (%22.63) iken deney grubunda olası etkileşim sayısı 858 (22x39) ve anlamlı bulunan etkileşim sayısı 160 (%18.65) olduğu bulunmuştur. Anlamlı bulunan etkileşimlerin t-değerlerinin sahip olduğu işarete göre FPF iki şekilde ifade edilmektedir. Buna göre t istatistiği eksi değer aldığında farklılaşan puanlayıcı katılığı, artı değer aldığında ise farklılaşan puanlayıcı cömertliği olarak ifade edilmektedir. Deney ve kontrol grubundaki puanlayıcıların anlamlı bulunan etkileşim türüne ilişkin frekans ve yüzdeleri Tablo 1'de verilmiştir.

Tablo 1. Öntest Ölçümlerine İlişkin pxb Etkileşiminde Anlamlı Bulunan Etkileşimlere Ait Frekans ve Yüzdeler

Grup	Farklılaşan Puanlayıcı Katılığı		Farklılaşan Puanlayıcı Cömertliği		Toplam	
	f	%	f	%	f	%
Deney	83	9.67	77	8.98	160	18.65
Kontrol	111	12.37	92	10.26	203	22.63

Tablo 1 incelendiğinde deney ve kontrol gruplarının farklılaşan puanlayıcı fonksiyonlarının ölçümlere karışma düzeylerinin birbirine yakın olduğu görülmektedir. Kontrol ve deney grubundaki puanlayıcıların farklılaşan puanlayıcı katılık ve cömertliklerinin istatistiksel anlamlılığı ÇYRM etkileşim analizinde elde edilen yanlılık büyüklüğü (bias size) değerleri kullanılarak test edilmiş ve analiz sonuçları tablo 2'de verilmiştir.

Tablo 2. Deney ve Kontrol Grubundaki Öntest Ölçümlerine Ilişkin Anlamlı Etkileşimlerin Farklılaşmasına Ilişkin Mann Whitney U Testinin Sonuçları

FPF Türü	**Grup**	**N**	**Sıra Ortalaması**	**Z**	**U**
FPK	Kontrol	111	90.88	-1.90	3872.00
	Deney	83	106.35		
FPC	Kontrol	92	87.55	-0.74	3307.00
	Deney	77	81.95		

Not: * p<0,05; FPK = Farklılaşan Puanlayıcı Katılığı, FPC = Farklılaşan Puanlayıcı Cömertliği

Tablo 2 incelendiğinde, deney ve kontrol grubundaki puanlayıcıların puanlayıcı eğitimi öncesi farklılaşan puanlayıcı fonksiyonlarının performans değerlendirme sürecinde ölçümlere karışma düzeylerinin istatistiksel olarak birbirine benzediği bulunmuştur (FPK için U = 3872.00; Z = -1.90 p > 0.05; FBC için U = 3307.00; Z = -0.74; p > 0.05).

Puanlayıcı Eğitimi Sonrası Deney ve Kontrol Grubundaki Puanlayıcıların FPF Durumlarının İncelenmesi

Deneysel işlem sonrası *puanlayıcı x öğrenci kompozisyonları* (pxb) etkileşimlerinin grup düzeyindeki istatistiksel göstergesine ilişkin kestirilen ki-kare değerinin anlamlı olduğu bulunmuştur ($\chi 2$(sd) = 4 084.90 (1755), p <0.05). Bu bulgu puanlayıcı eğitimi verilmiş olmasına rağmen puanlayıcıların performans değerlendirme sürecinde farklılaşan puanlayıcı fonksiyonunun ölçümlere karıştığını göstermektedir.

FPF'inin grup düzeyinde ölçümlere karıştığı belirlendiğinden bireysel düzeydeki istatistikler incelenmiştir. Bu bağlamda pxb etkileşimlerine ilişkin t istatistikleri incelenmiştir. Kontrol grubunun 897 olası etkileşiminden 163'ü (%18.17) anlamlı iken, deney grubunun 858 olası etkileşiminden 110'unun (%12.82) anlamlı olduğu bulunmuştur. Puanlayıcı eğitiminden sonra deney ve kontrol grubundaki puanlayıcıların performans değerlendirme sürecinde ölçümlere karışan farklılaşan puanlayıcı fonksiyonuna ilişkin frekans ve yüzde değerleri Tablo 3'te verilmiştir.

Tablo 3. Sontest Ölçümlerine Ilişkin pxb Etkileşiminde Anlamlı Bulunan Etkileşimlere Ait Frekans ve Yüzdeler

Grup	Farklılaşan Puanlayıcı Katılığı		Farklılaşan Puanlayıcı Cömertliği		Toplam	
	f	%	f	%	f	%
Deney	59	6.88	51	5.94	110	12.82
Kontrol	95	10.59	68	7.58	163	18.17

Tablo 3 incelendiğinde verilen puanlayıcı eğitimi sonrası deney ve kontrol gruplarındaki puanlayıcıların öğrenci kompozisyonlarını değerlendirme sürecinde FPF'in ölçümlere karışma düzeyinin farklılaştığı görülmektedir. Kontrol ve deney grubundaki puanlayıcıların farklılaşan puanlayıcı katılık ve cömertliklerinin istatistiksel anlamlılığı ÇYRM etkileşim analizinde elde edilen yanlılık büyüklüğü (bias size) değerleri kullanılarak test edilmiş ve analiz sonuçları tablo 4'de verilmiştir.

Tablo 4. Deney ve Kontrol Grubundaki Sontest Ölçümlerine İlişkin Anlamlı Etkileşimlerin Farklılaşmasına Ilişkin Mann Whitney U Testinin Sonuçları

FPF Türü	**Grup**	**N**	**Sıra Ortalaması**	**Z**	**U**	***p***	**d**
FPK	Kontrol	95	69.82	-2.72	2072.50	0.007*	0.22
	Deney	59	89.87				
FPC	Kontrol	68	56.21	-1.38	1476.50	0.167	--
	Deney	51	65.05				

Not: * p<0,05; FPK = Farklılaşan Puanlayıcı Katılığı, FPC = Farklılaşan Puanlayıcı Cömertliği

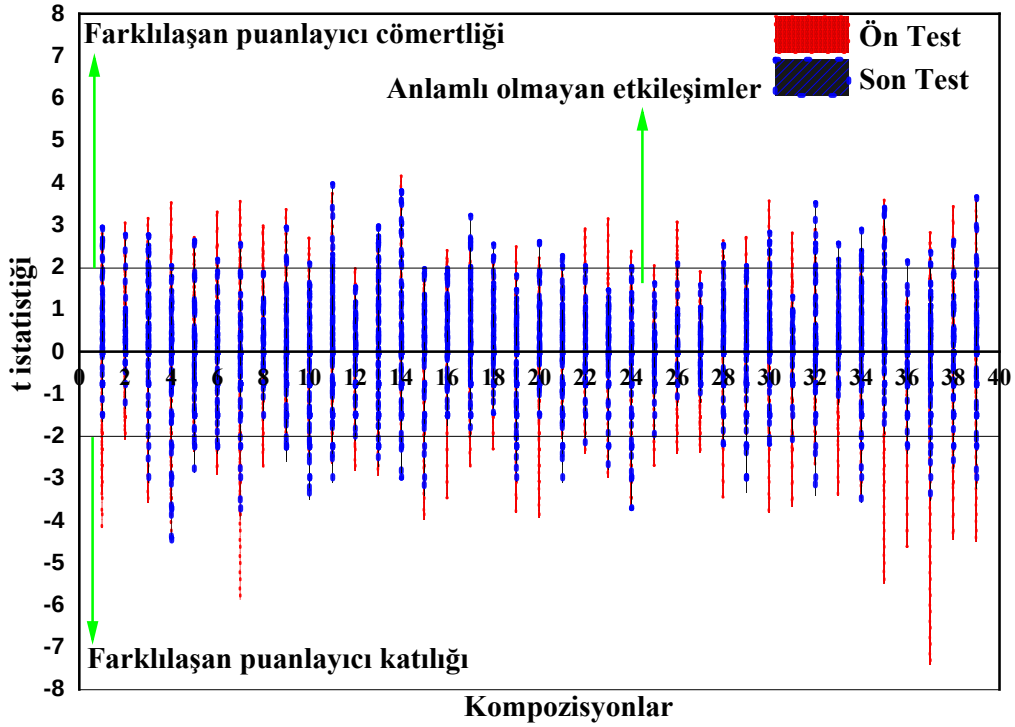
Tablo 4 incelendiğinde, puanlayıcı eğitimi sonrası performans değerlendirme sürecinde farklılaşan puanlayıcı katılığının ölçümlere karışma düzeyi gruplara göre istatistiksel olarak anlamlı çıkmış iken farklılaşan puanlayıcı cömertliğinin ölçümlere karışma düzeyi anlamsız bulunmuştur (FPK için U = 2072.50; Z = -2.72 $p < 0.05$; FBC için U = 1476.50; Z = -1,38; $p > 0.05$). Bu sonuca göre puanlayıcı eğitiminin farklılaşan puanlayıcı katılığı üzerinde küçük bir etkiye (r = 0.22) sahip olduğu, farklılaşan puanlayıcı cömertliği üzerinde ise herhangi bir etkiye sahip olmadığı bulunmuştur.

Puanlayıcı eğitiminin pxb etkileşimleri üzerindeki etkisini görmek amacıyla deney grubundaki puanlayıcıların ön ve son teste göre anlamlı etkileşim sayıları Tablo 5'te verilmiştir.

Tablo 5**.** Deney Grubundaki Puanlayıcılara İlişkin Anlamlı pxb Etkileşimleri

Test		P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11
Öntest	f	9	7	7	13	7	11	7	5	13	9	2
	%	23.1	18.0	18.0	33.3	18.0	28.2	18.0	12.8	33.3	23.1	5.1
Sontest	f	2	9	10	5	4	2	2	6	6	1	8
	%	5.1	23.1	25.6	12.8	10.3	5.1	5.1	15.4	15.4	2.6	20.5
		P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22
Öntest	f	3	9	7	13	3	5	7	5	4	6	9
	%	7.7	23.1	18.0	33.3	7.7	12.8	18.0	12.8	10.3	15.4	23.1
Sontest	f	4	5	8	8	2	2	6	8	3	3	9
	%	10.3	12.8	20.5	20.5	5.1	5.1	15.4	20.5	7.7	7.7	23.1

Tablo 5 incelendiğinde, puanlayıcı eğitimi sonrası öğrenci kompozisyonlarının değerlendirilmesi sürecinde; 14 puanlayıcının (1, 4, 5, 6, 7, 9, 10, 13, 15, 16, 17, 18, 20 ve 21 numaralı puanlayıcılar) anlamlı etkileşimlerinin azaldığı (eğitimden olumlu yönde etkilenme), 7 puanlayıcının (2, 3, 8, 11, 12, 14 ve 19 numaralı puanlayıcılar) anlamlı etkileşim sayılarının arttığı (eğitimden olumsuz etkilenme) ve 1 puanlayıcının (22 numaralı puanlayıcı) ise anlamlı etkileşim sayısının sabit kaldığı görülmektedir. Tablo 5'in daha anlaşılır olması amacıyla pxb etkileşimlerinin grafiksel gösterimi şekil 1'de verilmiştir.



Şekil 1. Deney Grubundaki Tüm Puanlayıcılar için pxb Etkileşimleri

Şekil 1 incelendiğinde, puanlayıcıların ön testini temsil eden kırmızı çizgilerin daha fazla ±2 aralığının dışında olduğu ve puanlayıcıların puanlayıcı eğitiminden sonraki puanlamalarını temsil eden mavi çizgilerin ise daha az ±2 aralığının dışında olduğu gözlemlenmektedir. Şekil 1'e göre bazı kompozisyonların diğer kompozisyonlara göre daha fazla puanlayıcı yanlılığına maruz kaldığı görülmektedir. Örneğin, puanlayıcılar diğer kompozisyonlara göre 37 numaralı kompozisyonun değerlendirilmesi sürecinde daha fazla katı davranmıştır. Ayrıca verilen puanlayıcı eğitiminin genel olarak puanlayıcı hataları üzerinde olumlu bir etki meydana getirdiği ve bunun sonucunda ölçümlerin geçerliğine katkı sağladığı söylenebilir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada ikinci dil akademik yazma becerisinin değerlendirilmesi sürecinde ölçümlere karışan FPF üzerinde puanlayıcı eğitiminin etkisinin araştırılması amaçlanmıştır. Bu bağlamda puanlayıcı eğitiminin öncesi ve sonrasına yönelik elde edilen bulgular incelenmiştir. Puanlayıcı eğitimi öncesi hem deney hem de kontrol grubunun öğrenci kompozisyonlarını değerlendirme sürecinde ölçümlere karışan FPF etkisinin benzer olduğu bulunmuştur. Hem grup düzeyinde hem de bireysel istatistiklerde benzer FPF etkilerinin olduğu bulunmuştur. Deney ve kontrol grubundaki pxb etkileşimlerinin yaklaşık beşte birinin FPF olduğu bulunmuştur. Alanyazın incelendiğinde, performans değerlendirme sürecinde ölçümlere FPF'in sıklıkla karıştığı görülmektedir (Liu ve Xie, 2014; Schaefer, 2008; Wesolowski ve diğerleri, 2015; Wolfe ve McVay, 2012). Öğrenci kompozisyonlarının değerlendirilmesi sürecinde ölçümlere karışan FPF iki şekilde ortaya çıkmıştır. Birincisi farklılaşan puanlayıcı katılığı iken ikincisi farklılaşan puanlayıcı cömertliği olmuştur. Araştırmada puanlayıcıların performans değerlendirme sürecinde farklılaşan puanlayıcı katılığını daha fazla gösterdikleri bulunmuştur. Alanyazın incelendiğinde performans değerlendirme sürecinde ölçümlere karışan FPF'in hem katılık hem de cömertlik davranışının bir kombinasyonu olduğu ve genellikle çok katı veya çok cömert puanlayıcılardan dolayı FPF'in ortaya çıktığı bulunmuştur (Kim ve diğerleri, 2012). Mevcut araştırmada katı puanlayıcıların daha fazla olduğu göz önüne alındığında farklılaşan puanlayıcı katılığının daha fazla olması sonuçları ve alanyazını doğrulamaktadır.

Puanlayıcı eğitiminden sonra öğrenci kompozisyonlarının değerlendirilmesi sürecinde FPF'in ölçümlere karışma düzeyi incelenmiş ve kontrol grubundaki değişim miktarı çok az iken deney grubunda önemli bir değişim olduğu tespit edilmiştir. Deney ve kontrol gruplarının puanlayıcı eğitimi öncesi FPF'in iki türünün ölçümlere karışma düzeyi istatistiksel olarak benzer iken, puanlayıcı eğitimi sonrası istatistiksel olarak farklılık göstermiştir. Farklılaşan puanlayıcı cömertliğinin deneysel işlemden etkilenmediği farklılaşan puanlayıcı katılığının etkilendiği bulunmuştur. Başka bir deyişle puanlayıcı eğitimi FPF'in farklılaşan puanlayıcı katılığı üzerinde etkili olmuştur. Bijani (2018), Fahim ve Bijani (2011), May (2008) ve Yan (2014) tarafından yapılan araştırmalar incelendiğinde, puanlayıcı eğitiminin FPF üzerinde etkili olduğu bulunmuştur. Van Dyke (2008) tarafından yapılan araştırmada ise performans değerlendirme sürecinde farklılaşan puanlayıcı cömertliğin ölçümlere karıştığı farklılaşan puanlayıcı katılığın ise karışmadığı bulunmuştur. Mevcut araştırmanın Van Dyke (2008) tarafından yapılan araştırma ile desteklenmemesinin temel nedenlerinden birinci puanlayıcıların farklı gruplardan oluşması ikincisi ise değerlendirilmesi yapılan performansın farklı olmasından kaynaklandığı düşünülmektedir.

Öğrencilerin ikinci dildeki akademik yazma becerilerinin değerlendirilmesi sürecinde ölçümlere karışan FPF üzerinde puanlayıcı eğitiminin etkisinin araştırıldığı bu araştırmada elde edilen sonuçlar aşağıda verildiği şekilde özetlenebilir;

- Öğrenci kompozisyonlarının değerlendirilmesi sürecinde FPF'in ölçümlere karıştığı ve yaklaşık olarak pxb etkileşimlerinin beşte birini oluşturduğu bulunmuştur.
- Deney ve kontrol grubundaki puanlayıcıların puanlayıcı eğitimi öncesi benzer FPF sergiledikleri bulunmuştur.
- Puanlayıcı eğitimi FPF'in farklılaşan puanlayıcı katılığı türü üzerinde etkili olduğu ve puanlayıcı eğitiminin FPF üzerinde küçük bir etki büyüklüğüne sahip olduğu bulunmuştur.

Araştırma kapsamında elde edilen sonuçlardan hareketle gelecekte yapılacak çalışmalar ve araştırmacılar için bazı önerilerde bulunulmuştur;

- Mevcut araştırmada iki farklı puanlayıcı eğitimi deseni birleştirilerek verilmiştir. Alanyazında birçok farklı puanlayıcı eğitimi deseninin olduğu göz önüne alındığında farklı kombinasyonlar yapılarak puanlayıcı eğitimlerinin FPF üzerindeki etkileri incelenebilir.
- Bu araştırmada kalabalık bir deney grubu kullanılmış olup alanyazında daha küçük (n =5-6) gruplara verilen eğitimin etkili olduğu bu bağlamda gelecek çalışmalarda küçük grupların kullanılması faydalı olabilir.
- Puanlayıcı eğitiminin FPF üzerinde etkili olması yerleşme ve seçme sınavlarında kullanılan performans değerlendirme sürecinde puanlayıcıların eğitilmesinde kullanılarak ölçümlerin geçerliğine ve güvenirliğine katkı sağlanabilir.

KAYNAKÇA

Aryadoust, V. (2016). Understanding the growth of ESL paragraph writing skills and its relationships with linguistic features. *Educational Psychology*, *36*(10), 1742-1770. https://doi.org/10.1080/01443410.2014.950946

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. https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1603 sayfasından erişilmiştir.

Baştürk, M. (2012). İkinci dil öğrenme algılarının belirlenmesi: Balıkesir örneği. *Balikesir University Journal of Social Sciences Institute, 15*(28-1), 251-270. http://dspace.balikesir.edu.tr/xmlui/handle/20.500.12462/4594 sayfasından erişilmiştir.

Bayat, N. (2014). Öğretmen adaylarının eleştirel düşünme düzeyleri ile akademik yazma başarıları arasındaki ilişki. *Eğitim ve Bilim*, *39*(173), 155-168. http://eb.ted.org.tr/index.php/EB/article/view/2333 sayfasından erişilmiştir.

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

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, *6*(2), 205-212. https://journals.aom.org/doi/abs/10.5465/amr.1981.4287782 sayfasından erişilmiştir.

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

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL students. *Journal of Second Language Writing*, *14*, 191–205. https://doi.org/10.1016/j.jslw.2005.08.001

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

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.* (Doktora Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

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

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. New York: 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*. Ankara: Pegem Akademi Yayıncılık.

Carter, C., Bishop, J. L., & Kravits, S. L. (2002). *Keys to college studying: becoming a lifelong learner*. New Jersey: Printice Hall.

Çekici, Y. E. (2018). Türkçe'nin yabancı dil olarak öğretiminde kullanılan ders kitaplarında yazma görevleri: Yedi iklim ve İstanbul üzerine karşılaştırmalı bir inceleme. *Gaziantep Üniversitesi Eğitim Bilimleri Dergisi*, *2*(1), 1-10. http://dergipark.gov.tr/http-dergipark-gov-tr-journal-1517-dashboard/issue/36422/367409 sayfasından erişilmiştir.

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

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*(1), 1–8. https://doi.org/10.1080/15434303.2011.622016

Cumming, A. (2014). Assessing integrated skills. In A. Kunnan (Vol. Ed.), *The companion to language assessment: Vol. 1*, (pp. 216–229). Oxford, United Kingdom: Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla131

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., & Frisbie, D.A. (1991). *Essentials of educational measurement*. New Jersey: 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*. Frankfurt: 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 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*, i-60. https://doi.org/10.1002/j.2333-8504.2003.tb01893.x

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahway, NJ: Lawrence Erlbaum Associates

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. http://www.ijlt.ir/portal/files/401-2011-01-01.pdf sayfasından erişilmiştir.

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://jalt-publications.org/files/pdf-article/jj2012a-art4.pdf sayfasından erişilmiştir.

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. https://pdfs.semanticscholar.org/dd21/ba5683dde8b616374876b0c53da376c10ca9.pdf sayfasından erişilmiştir.

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

Gillet, A., Hammond, A. & Martala, M. (2009). *Successful academic writing*. New York: Pearson Longman.

Göçer, A. (2010). Türkçe öğretiminde yazma eğitimi. *Uluslararası Sosyal Araştırmalar Dergisi*, *12* (3), 178-195. http://www.sosyalarastirmalar.com/cilt3/sayi12pdf/gocer_ali.pdf sayfasından erişilmiştir.

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.* New Jersey: Prentice-Hall Press

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological bulletin*, *103*(2), 265-275. https://doi.org/10.1037/0033-2909.103.2.265

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

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*. Harlow: Pearson Education.

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

IELTS (t.y). *Prepare for IELTS*. https://takeielts.britishcouncil.org/prepare-test/free-sample-tests/writing-sample-test-1-academic/writing-task-2 sayfasından erişilmiştir.

İlhan, M. (2015). *Standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi*. (Doktora Tezi). https://tez.yok.gov.tr sayfasından erişilmiştir.

İlhan, M., & Çetin, B. (2014). Performans değerlendirmeye karışan puanlayıcı etkilerini azaltmanın yollarından biri olarak puanlayıcı eğitimleri: Kuramsal bir analiz. *Journal of European Education*, *4*(2), 29-38. https://doi.org/10.18656/jee.77087

Jin, K. Y., & Wang, W. C. (2017). Assessment of differential rater functioning in latent classes with new mixture facets models. *Multivariate behavioral research*, *52*(3), 391-402. https://doi.org/10.1080/00273171.2017.1299615

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

Kassim, N. L. A (2007). *Exploring rater judging behaviour using the many-facet Rasch model.* Paper Presented in the Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2), Universiti Utara, Malaysia. http://repo.uum.edu.my/3212/ sayfasından erişilmiştir.

Kassim, N. L. A. (2011). Judging behaviour and rater errors: an application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, *11*(3), 179-197. http://ejournals.ukm.my/gema/article/view/49 sayfasından erişilmiştir.

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

Kim, Y.K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model* (Doktora Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

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. https://eric.ed.gov/?id=EJ920513 sayfasından erişilmiştir.

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

Kutlu, Ö., Doğan, C.D., & Karaya, İ. (2014). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme*. Ankara: Pegem Akademi Yayıncılık.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, *28*(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J. M. (1993). Rasch-based generalizability theory. *Rasch Measurement Transaction*, *7*(1), 283-284. https://www.rasch.org/rmt/rmt71h.htm sayfasından erişilmiştir.

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

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. *Objective measurement: Theory into practice*, 3, 85-98.https://files.eric.ed.gov/fulltext/ED364573.pdf sayfasından erişilmiştir.

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

Liu, J., & Xie, L. (2014). Examining rater effects in a WDCT pragmatics test. *Iranian Journal of Language Testing*, *4*(1), 50-65. https://cdn.ov2.com/content/ijlte_1_ov2_com/wp-content_138/uploads/2019/07/422-2014-4-1.pdf sayfasından erişilmiştir.

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*. Mahwah, NJ: Erlbaum.

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

Moore, B.B. (2009). *Consideration of rater effects and rater design via signal detection theory*. (Doktora Tezi). http://www.proquest.com/ sayfasından erişilmiştir.

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?*. http://pareonline.net/htm/v7n3.htm sayfasından erişilmiştir.

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. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422. http://psycnet.apa.org/record/2003-09517-007 sayfasından erişilmiştir.

Oosterhof, A. (2003). *Developing and using classroom assessments*. New Jersey: 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

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, 94(1), 31-37. http://peterliljedahl.com/wp-content/uploads/Myth-of-Objectivity2.pdf sayfasından erişilmiştir.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Selden, S., Sherrier, T., & Wooters, R. (2012). Experimental study comparing a traditional approach to performance appraisal training to a whole-brain training method at CB Fleet Laboratories. *Human Resource Development Quarterly*, *23*(1), 9-34. https://doi.org/10.1002/hrdq.21123

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). New York: 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

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes*, *8*, 207-223. https://doi.org/10.1016/j.jeap.2009.03.001

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

Van Dyke, N. (2008). Self-and peer-assessment disparities in university ranking schemes. *Higher Education in Europe*, *33*(2/3), 285-293. https://doi.org/10.1080/03797720802254114

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: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wesolowski, B. C., Wind, S. A., & Engelhard Jr, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147-170. https://doi.org/10.1177/1029864915589014

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

Wind, S. A., & Guo, W. (2019). Exploring the combined effects of rater misfit and differential rater functioning in performance assessments. *Educational and psychological measurement*, *79*(5), 962-987. https://doi.org/10.1177/0013164419834613

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

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31-37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*(4), 501-527. https://doi.org/10.1177/0265532214536171

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

Zwiers, J. (2008). *Building academic language: Essential practices for content classrooms*. San Francisco: Jossey-Bass.

Ek A. Yazma Görevi Örneği

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.

Ek B. Akademik Yazma Becerileri İçin Analitik Dereceli Puanlama Anahtarı

	ORGANIZATION					CONTENT	
Point	Introduction-Body-Conclusion	Thesis Statement	Topic Sentence	Supporting Sentences	Appropriate Length	Topic Relevance	Idea Development
4	The organization of introduction, body, and conclusion paragraphs is *highly* appropriate to written genre.	Thesis statement is *noticeably* given in introduction paragraph. It *comprehensively* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *comprehensively* addresses and supports the specific idea(s) given in thesis statement. It *extensively* demonstrates the main idea of the paragraph.	Supporting sentences *comprehensively* illustrate the main idea given in topic sentence.	There are *at least 250 words* in written text. It is constructed with *appropriate length*.	Written text is *highly* relevant to assigned topic in task. It *comprehensively* addresses all parts of the task.	*Extensive* details are provided to develop, support and illustrate information or ideas presented in written text.
3	The organization of introduction, body, and conclusion paragraphs is *largely* appropriate to written genre.	Thesis statement is *evidently* given in introduction paragraph. It *mostly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *mostly* addresses and supports the specific idea(s) given in thesis statement. It *largely* demonstrates the main idea of the paragraph.	Supporting sentences *adequately* illustrate the main idea given in topic sentence.	Text length is between *200 and 249 words*. It is *slightly* shorter than required length.	Written text is *mostly* relevant to assigned topic in task. It *adequately* addresses the basic parts of the task.	*Adequate* details are provided to develop, support and illustrate information or ideas presented in written text.
2	The organization of introduction, body, and conclusion paragraphs is *moderately* appropriate to written genre.	Thesis statement is *less explicitly* given in introduction paragraph. It *moderately* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *moderately* addresses and supports the specific idea(s) given in thesis statement. It demonstrates the main idea of the paragraph in *some respects.*	Supporting sentences *moderately* illustrate the main idea given in topic sentence.	Text length is between 150 *and 199 words*. It is *seemingly* shorter than required length.	Written text is *moderately* relevant to assigned topic in task. It *partially* addresses the basic parts of task.	*Basic* details are provided to develop, support and illustrate information or ideas presented in written text.
1	There is *inadequate* organization of introduction, body, and conclusion paragraphs in the written text.	Thesis statement is *vaguely* given in introduction paragraph. It *slightly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *partially* addresses and supports the specific idea(s) given in thesis statement. It *slightly* demonstrates the main idea of the paragraph.	Supporting sentences *partially* illustrate the main idea given in topic sentence.	Text length is between 100 *and 149 words*. It is *considerably* shorter than required length.	Written text is *slightly* relevant to assigned topic in task. It lacks addressing the basic parts of the task.	*Some details are* provided but they are not enough to develop, support and illustrate information or ideas presented in written text.
0	Written text lacks organization of introduction, body and conclusion paragraphs.	Thesis statement is not given in introduction paragraph or it does not include any specific idea(s) to be elaborated in the written text.	Topic sentence is not included in written text, or it does not address the thesis statement or demonstrate the main idea of the paragraph.	Written text does not include supporting sentences or they do not illustrate the main idea given in topic sentence.	Text length is *below 99 words.* It does not meet the requirement of appropriate length.	Written text is irrelevant to assigned topic in task. It fails to address the task adequately.	Information or ideas are not *thoroughly* developed, supported or illustrated in written text.

	COHERENCE	COHESION	GRAMMAR		VOCABULARY		MECHANICS	
Point	Coherence	Linking	Accuracy of Grammatical Forms	Syntactic Complexity	Word Choice	Lexical Range	Spelling	Punctuation
4	Information or ideas sequenced in paragraphs are *highly* consistent. There is a *considerably* logical progression between sentences in written text.	A *wide* range of cohesive devices used to connect ideas in written text provides a smooth transition between sentences.	All grammatical forms are *accurately* used in written text. The communication is *successfully* established.	Complex and sophisticated sentences are *extensively* used in written text in which syntactic structures are *highly* diverse.	All the words and phrases are *appropriately* used. The intended meaning is *clearly* conveyed in written text.	There is a *wide range* of vocabulary used in written text which includes *highly* sophisticated words and phrases.	All the needed spelling rules are *accurately* used in written text.	All the needed punctuation rules are *accurately* used in written text.
3	Information or ideas sequenced in paragraphs are *mostly* consistent. There is an *adequately* logical progression between sentences in written text.	An *adequate* range of cohesive devices used to connect ideas in written text provides an easy transition between sentences.	The use of the grammatical forms is *mostly accurate* in the written text. There are *few grammatical errors* which do not impede communication.	Complex and sophisticated sentences are *widely* used in written text in which syntactic structures are *adequately* diverse.	The use of words and phrases is *mostly appropriate*. There are *few* misused words or phrases which cannot obscure the intended meaning.	There is an *adequate range* of vocabulary used in written text which includes *largely* sophisticated words and phrases.	All the needed spelling rules are *mostly accurate* in written text but there are *few errors* which violate these rules.	All the needed punctuation rules are *mostly accurate* in written text but there are few errors which violate these rules.
2	Information or ideas sequenced in paragraphs are *moderately* consistent but there are some inconsistencies which *partially* interrupt logical progression between sentences.	The use of cohesive devices *at basic level* to connect ideas in written text provides a complete transition between sentences.	It is attempted to use the grammatical forms accurately in written text but there are *occasional grammatical errors* which slightly impede communication.	Complex and sophisticated sentences are *moderately* used in written text in which syntactic structures are *partially* diverse.	It is attempted to use the words and phrases appropriately but there are *occasionally* misused words or phrases which *slightly* obscure the intended meaning.	The *basic* vocabulary is used in written text which includes *moderately* sophisticated words and phrases.	It is intended to use the needed spelling rules *accurately* in written text but there are *occasional errors* which violate these rules.	It is intended to use the needed punctuation rules *accurately* in written text but there are *occasional errors* which violate these rules.
1	Paragraphs are constructed with *slightly* consistent information or ideas which interrupt logical progression and sequence between sentences.	A *limited* range of cohesive devices used to connect ideas in written text makes transition between sentences fragmentary.	The use of the grammatical forms is *generally inaccurate* in written text. There are *frequent grammatical errors* which largely impede communication.	Complex and sophisticated sentences are *slightly* used in written text in which syntactic structures are diverse to some extent.	The use of words and phrases is *generally inappropriate*. There are *frequently* misused words or phrases which *largely* obscure the intended meaning.	There is a *limited range* of vocabulary used in written text which includes *slightly* sophisticated words and phrases.	The use of the needed spelling rules is *largely* inaccurate. There are *frequent errors* which violate these rules.	The use of the needed punctuation rules is *largely* inaccurate. There are *frequent errors* which violate these rules.
0	Written text lacks consistency and logical progression between sentences.	There is an *inadequate* use of cohesive devices in written text which lacks transition between sentences.	The use of grammatical forms is *completely inaccurate* in the written text. This causes a breakdown in communication.	Written text lacks sentential complexity, sophistication and syntactic variety.	The use of vocabulary is completely inappropriate in written text. The intended message is obscured.	A repetitive vocabulary is largely used in written text which lacks sophistication.	All the needed spelling rules are *inaccurately* used in written text.	All the needed punctuation rules are *inaccurately* used in written text.