Id: 900293

Participatory Educational Research (PER)
Vol. 8(4), pp. 239-252, December 2021
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.21.88.8.4

Examining the Differential Rater Functioning in the Process of Assessing Writing Skills of Middle School 7th Grade Students

Aslıhan Erman Aslanoğlu *
Faculty of Education, Ufuk University, Turkey
ORCID: 0000-0002-1364-7386

Mehmet Şata
Faculty of Education, Agri Ibrahim Cecen University,Turkey
ORCID: 0000-0003-2683-4997

Article history

Received:
02.03.2021

Received in revised form:
16.04.2021

Accepted:
17.04.2021

Key words:
Writing assessment, Differential rater functioning, Many-facet, Rasch measurement

When students present writing tasks that require higher order thinking skills to work, one of the most important problems is scoring these writing tasks objectively. The fact that raters give scores below or above their performance based on several environmental factors affects the consistency of the measurements. Inconsistencies in scoring negatively affect the validity and reliability of student performance and cause the scores obtained to be questioned. In regard to the validity and reliability of these measurements, it is significant to identify the rater behavior and correct the sources of error. This study aims to analyze the differential rater functioning (DRF), which is one of the problematic rater behaviors, in evaluating compositions written by middle school 7th-grade students within the scope of the Turkish course. 86 students attending a public school were participated the study. Students' compositions were rated using an analytical rubric by 8 teachers from different institutions. In this correlational research, the many facet Rasch model was used, and five variables including students, raters' and, students' gender, students' qualification, and evaluation criteria were examined. it was examined whether the raters show DRF on an individual and group basis based on the dual interaction analysis, including the gender of the student x rater and the student's competence x rater. The findings have revealed that DRF at the group level does not interfere with the measurements, while the individual level DRF is involved in the measurements. It was determined that the level of DRF mixing in the measurements of successful students was the lowest. Especially rigid and lenient raters were found to show DRF. In the present study, it was observed that the raters showing DRF was also the most lenient raters, while these raters did not show DRF in terms of the gender of the student.

Introduction

Writing can be defined as writing down of the information that individuals organize in their minds via putting this organized information on paper (Englert & Mariage, 2003). In other words, writing can be seen as a way for individuals to express themselves by organizing their

* Correspondency: aslihanerman@yahoo.com

knowledge, feelings, and thoughts. In this respect, writing is an important tool for individuals to express themselves. Researchers state that in the teaching process, effective use of writing better enables students to express the newly learned information and their thoughts, which improves their written communication skills as well as their academic success in other courses (Erhardt & Meade, 2005; Lam, Au, Leung, & Li Tsang., 2011).

Students are expected to perform the processes of designing, organizing, drafting, editing, and editing effectively while they are writing (Englert, 1991). In this respect, the thinking method that students use while creating a written product is the cognitive aspect of writing, and the method of checking when producing a written product are the metacognitive aspects of writing (Collins, 2000). Writing is a process that requires individuals to use higher order thinking skills defined as associating these skills with individual characteristics by using more than one skill simultaneously (Marzono, 2001).

When students show a performance that requires the use of higher order thinking skills like mental skills when writing, one of the most important issues is the measurement of these performances. Although there are educational outputs to improve writing skills in programs teaching Turkish as the mother tongue, there is no standard measurement method to evaluate the extent to which students have this basic skill. However, performance assessment approaches, which are considered important by researchers in measuring higher order thinking skills, are also used in measuring, and evaluating writing skills. Performance evaluation approaches try to measure to what extent students are good at utilising the basic knowledge and new-learned skills during task performance which require more complex and higher order thinking skills in realistic conditions (Erman Aslanoğlu & Kutlu, 2003). A rubric is used to ensure the reliability of the measurements made in performance evaluation (Russell & Airasian, 2011). Graded scoring keys are scoring tools that contain the list of criteria for a task and the degree of qualifications related to these criteria (Goodrich, 1997). Although the graded scoring key is used to ensure the validity and reliability of the scores in the performance evaluation process, there may be unwanted results in the measurement results of the students, in other words, possible mistakes may get involved in this process. The most important source of error in performance evaluation is caused by raters.

Errors in rater decisions, in other words, rater effects, can affect the accuracy of assigned ratings. Rater-driven factors that negatively affect validity and reliability are called rater effects (Farrokhi, Esfandiari, & Vaez Dalili, 2011). Although there are many errors due to raters in performance evaluations, the most common mistakes with rater effect in the literature are as follows: rater severity and leniency are considered as the halo effect, central tendency behavior, range restriction, and differential rater rigidity and generosity which might also be called rater bias or Differential Rater Function (DRF) (Myford & Wolfe, 2004).

The differential rater function (DRF) originating from the rater, which is also the subject of the present study and accepted as one of the important sources of error, scores become higher or lower for some individuals than others depending on the various characteristics of the rater, such as gender, age, cultural factors, while performing the evaluation process, which can be defined as a tendency to give different points to particular students regardless of the writing skill (Wesolowski, Wind, & Engelhard, 2015). For example, a rater can score male candidates' writing tasks more severity. Thus, DRF refers to a situation where the probability of students with the same basic ability level to be rated at the same level by raters due to their group membership, is not equal. As an example, an incorrect (bias) rater prefers or dislikes a particular group of students compared to another group when grading students' writing skills. When

writing products are scored by raters who know the participants or can predict the participant's gender or ethnicity, rater error may occur. If participants tend to score higher than their race's raters in an exam they take, these participants may have an unfair advantage (Schaefer, 2008). DRF mostly occurs when group membership is known. However, DRF can also occur in the case of unknown group membership. In other words, DRF seems possible when the group membership is not known and a structure that can be predicted from the data is formed (Jin & Wang, 2017). For example, in an activity that measures the writing skill, raters can give higher scores to students with good handwriting or looking at the information about the gender of the student from the name written on the paper. Jin and Wang (2017) tried to avoid this limitation by developing a new facet model to determine DRF when group membership was unknown. In both cases, raters give scores below or above their performance based on several environmental factors that affect the consistency of the measurements (Engelhard, 2008; Myford & Wolfe, 2009; Tamanini, 2008). Inconsistencies in scoring negatively affect the validity and reliability of student performance and cause questioning of the scores obtained (Eckes, 2009; Schaefer, 2008). Previous studies on DRF have focused on rater bias by analysing such manifested variables as the examinee's gender and race/ethnicity to determine the subgroups (Hoyt, 2000; Wesolowski, Wind, & Engelhard, 2015).

Studies have unearthed that rater effects are widespread and their effects can be mitigated by rater training and monitoring efforts (Feldman, Lazzara, Vanderbilt & DiazGranados, 2012; Hauenstein, & McCusker, 2017; Şata, 2019), and have demonstrated that many facet Rasch models (MFRM) can foster the detectability and understandability of the nature of these effects (Engelhard & Myford, 2003; Kim, Park & Kang, 2012; Wolfe, Chiu, & Myford, 2000).

MFRM expands the basic Rasch model by allowing researchers to add the facet of judge severity to person ability and item difficulty (Bond & Fox, 2015). In other words, MFRM measurement model considers all the sources of variability that are thought to affect the test score of individuals and provides a statistical approach that reveals the interaction of these various sources (Haiyang, 2010). The MFRM includes at least two sources of variability, and the measurements of these facets can be analyzed at once and independently of the sample. Besides, with the help of this model, individual and group-level evaluations of facets can be made (Linacre, 2018).

As Engelhard (1994) states, MFRM improves the fairness and objectiveness of the measurement of writing skill because writing ability may either be over or underestimated even due to the raw scores alone if students of a similar level are rated by raters with differing severity. MFRM adjusts for rater variability, which means that it presents a more accurate understanding of the skill which is being evaluated.

Literature Review

Utilizing MFRM analysis, studies focusing on determining rater bias seek to reveal the unexpected interactions between the decisions of the evaluator and the performance of the test takers. Having analyzed rater mediated writing performance assessments under the framework of the Rasch measurement (MFRM) approach, several researchers have revealed that some raters display systematic leniency or severity in combination with differential rater functioning (DRF) depending on rater, student, and/or test characteristics. In this account, studies related to DRF analysis mostly focused on the variables such as student gender, age race/ethnicity, language background, experience, having a rating training or not, and bias analysis of the topic type and rating criteria as well as the related test-related features. For example, Du, Wright, and

Brown (1996), in their study with university students through MFRM, determined that some raters showed differing rater bias according to the subject type affecting the writing skill scores of the students. Apart from this, it was observed that students at different ages and with different genders show different performances according to the subject type. As a result of the study conducted by Topaş (2020) with 51 students at the secondary school level and 11 raters, it was revealed that the raters showed generosity behaviour that differed according to the subject type, but it was observed that they did not exhibit DRF according to gender. In another study, Johnson and Lim (2009) found out as a result of their DRF analysis that the language background of the raters (native and non-native English speakers) had a very low interaction with the scores of the students, while some researchers observed that the native speakers evaluated the student scores more rigidly as a result of the bias analysis they applied (e.g. Engelhard & Myford, 2003; Kondo-Brown, 2002; Shi, 2001). In another study conducted by Kondo-Brown (2002), the compositions written by 284 students within the context of Japanese as a second language (L2) were evaluated by three teachers. They observed that the raters scored certain candidates and the criteria of the scoring key in a more tolerant or rigid way and that each rater's rater error model was different. Apart from this, they determined that the highest percentage of biased rater x candidate interactions were among the candidates with the highest or lowest ability. Engelhard and Myford (2003) examined the faculty consultants' rating behaviour while they are evaluating the essays written for the 1999 Advanced Placement English Literature and Composition Exam, and they discovered that some raters show DRF related to student gender, student race/ethnicity, or student best language. Additionally, Eckes (2005), working on the Test of German as a Foreign Language (TestDaf), investigated the severity and evaluative error/interaction of the evaluator in terms of test-takers, rubric's grading criteria, and gender. Significant rater bias was found among rater and test-takers, and rater and grading criteria, but no rater bias by gender.

There is no doubt that the criteria of the scoring key play an important role in studies where students' writing skills are evaluated. A series of studies on DRF show that especially raters with different backgrounds (native/non-native) show bias according to the grammatical criterion of the scoring key, accordingly, the grammar criterion is usually the criterion scored most biased by raters (e.g., native/non-native) (Eckes, 2005, 2008; Kondo-Brown, 2002; Mc Namara, 1996; Schafer, 2008). Schaefer (2008) also studied a combination of rater leniency/severity and DRF, especially in his research where he examined rater effects in an analysis of 40 essays composed by EFL students. In addition to leniency/severity effects, Schaefer's observation also informs that some raters show DRF in relation with the domains in the analytic scoring rubric, as well. In this sense, Schaefer (2008) determined bias patterns through his claim that raters were more tolerant with respect to accuracy in grammar when compared to their rating of other elements in writing such as organization. In Schaefer's research, it is clearly shown that "some raters also rated higher ability writers more severely and lower ability writers more leniently than expected".

Some researchers, using MFRM, reported the experience of raters in the context of rater bias. These studies found out that inexperienced raters display more strict behaviour than the experienced ones in evaluating writing performance (Choi, 2002; McNamara, 1996; Shi, 2001).

Finally, it was found out that some of the DRF studies on writing performance also consider self, peer, and teacher scores. For example, Farrokhi, Esfandiari, and Schaefer (2012) examined rater severity/generosity behaviour that differed in peer, self, and teacher evaluations made using an analytical rubric about English texts written by 188 students in Iran. The results of the study uncovered that the raters who made self-assessment and teacher evaluations made stricter

evaluations compared to the peer-assessment raters when assessing the highest and lowest talented students. Another researcher, Matsuno (2009), used MFRM for 91 students and 4 teacher evaluators to investigate how self and peer assessments work compared to teacher evaluations in college writing classes in Japan. As a result of the bias analysis he conducted to unearth the evaluator-author interactions, he found that the self-evaluation raters evaluated themselves more strictly. Moreover, this study found that highly successful authors did not rate their peers rigidly, while low-achieving authors often rated their peers more strictly. Finally, the most generous or strict raters among teacher raters often showed DRF.

The Present Study

In cases where more than one rater is used in the assessment of writing skills that require the introduction of higher order thinking skills, there may be inconsistencies between the scores and the errors resulting from these inconsistencies reduce the validity and reliability of the measurements. In terms of the validity and reliability of these measurements, it is important to determine the rater behaviour and correct the sources of error. In this account, among the rater behaviours in the current study; differential rater severity/leniency behaviour (bias interaction) was analyzed by MFRM. DRF is problematic in terms of student scores because when raters show this effect, student scores are not comparable across subgroups. It is thought that determining rater errors that may arise in the evaluation of higher order thinking skills such as writing skills with the MFRM approach will provide useful information to determine, understand and correct the nature of rater-induced errors. In addition to this, there is no study in native Turkish at the national level within the scope of the current research. In DRF studies conducted using the MFRM approach at the international level, it has been observed that English is generally considered within the scope of the second language (ESL). Determining rater bias and taking necessary precautions are considered important in measuring Turkish writing as the mother tongue.

The general purpose of the current study is to search the effect of the differential rater function which is one of the rater errors that is involved in the measurements in the process of evaluating middle school 7th-grade students' academic writing skills in their mother tongue, as a result of the interactions of rater x student qualities. In this regard, both rater mismatches (such as severity and leniency) and the level of interference in the measurements in the process of evaluating the writing performance of the differing rater function were examined on an individual and group basis according to gender and ability level variables.

For this purpose, answers to the following questions were sought in the study:

(1) Do the raters score the students according to their academic writing skill (unsuccessful, moderate, successful) and show the differential rater function at the group and individual level?

(1) Do the raters show the differential rater function at the group and individual level while scoring students according to their gender (female, male)?

Method

Setting and Participants

The correlational research model was used in the study since the relationship between teacher scores and student characteristics was aimed to be investigated in evaluating the academic writing skills of middle school students. In the present study, the many facet Rasch

model was used, and a completely crossed pattern was utilized due to crossing all the variables. In this study, there are five variables: students, raters, students' gender, students' competence status, and evaluation criteria. This study focuses on the surface interactions rather than individual variables. In this context, the focus of this study is on binary interaction analysis, with the students' gender x rater and the student's competence x rater.

Within the scope of the research, there are two types of participants, namely raters and students. Socio-demographic information about the raters in the study is given in Table 1.

Table 1. Frequency and percentages of the raters regarding their socio-demographic information

Variable	Level of Variable	Frequency	Percentage
Gender	Female	4	50.00
	Male	4	50.00
School Type	State	4	50.00
	Private	4	50.00
Professional Seniority	1 to 5 years	3	37.50
	6 to 10 years	3	37.50
	11 years and above	2	25.00
	Total	8	100

In Table 1, it is seen that the raters are equal in number according to gender and school type, and they are close to each other regarding professional seniority. Socio-demographic information about the students is given in Table 2.

Table 2. Frequency and percentage of students' socio-demographic information

Variable	Level of Variable	Frequency	Percentage
Gender	Female	47	54.65
	Male	39	45.35
Proficiency/Skill	Unsuccessful	31	36.05
	Moderate	17	19.77
	Successful	38	44.18
	Total	86	100

Table 2 illustrates that the distribution of students according to gender is close to each other and that the moderate-level group is less than the other two groups according to academic writing competencies.

Data Collection

To evaluate the academic writing skills of secondary school students, data were collected using an analytical rubric developed by the researchers. In the process of determining the criteria of a rubric, firstly, the opinions of three field experts and five teachers in the field of academic writing and the relevant literature were reviewed. Later, one of the compositions written by the students was scored by the teachers and the incomprehensible places were determined and rearranged. After all these stages, an analytical graded scoring key with six criteria and a four-point rating was developed.

Exploratory factor analysis and McDonald ω coefficient were used to provide evidence for the reliability of the measurements obtained from the developed analytical rubric and the validity of the inferences made based on these measurements. It was investigated whether the assumptions of the examined exploratory factor analysis were met and seen that the necessary assumptions were met (KMO value for the relevant data set was 0.866, Bartlett's test of

sphericity was significant; all criteria of the graded scoring key were normally distributed; no loss or the extreme value was found).

EFA was conducted to provide evidence for the construct validity of the analytical rubric (AR) developed to evaluate academic writing skills. During EFA, the average of the scores given by eight raters of the essays written by 86 students was taken. As a result of the data analysis, it was found that AR was collected under a single factor and explained 83.86% of the change in student achievement (Factor load of each criterion in the measurement tool was as follows; 0.856; 0.866; 0.943; 0.952 0.960 and 0.913). The fact that each criterion of AR has a factor load of 0.80 and above is an indication that the measurements have high discrimination and validity. The scattering diagram obtained as a result of AFA is given in Figure 1.

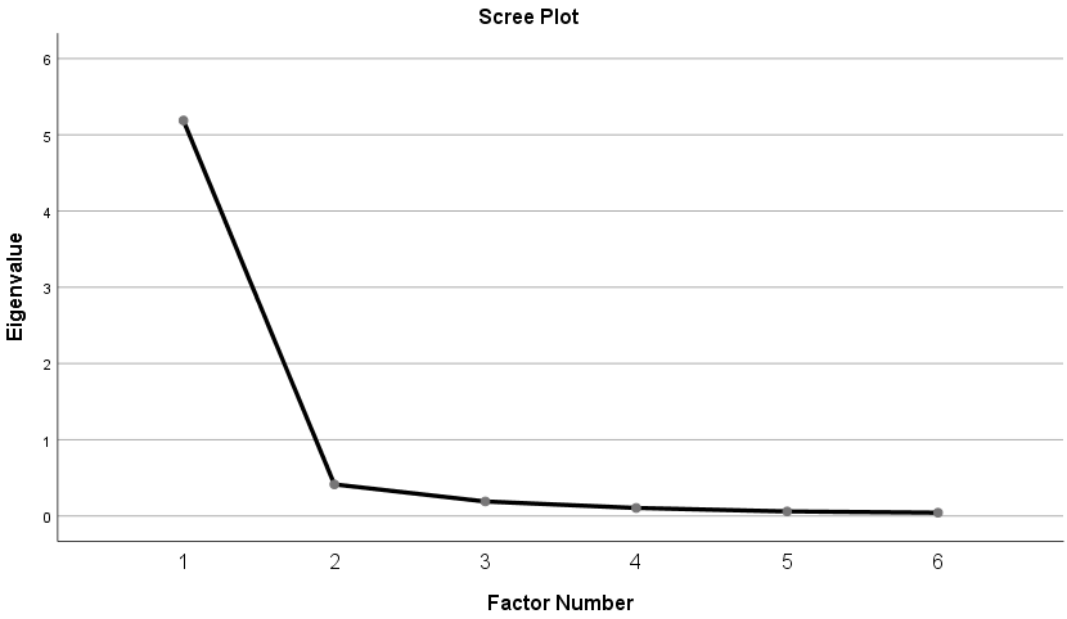

Figure 1. Scatter diagram for AR.

Looking at the ratio between the dominant factor in Figure 1 and the second highest factor, it is seen that it is approximately five times higher. In this context, the fact that all criteria have a high factor load under a single factor and that the pressure factor has a greater value than other factors is considered as evidence that the measuring tool has a one-dimensional structure.

The reliability coefficient (ω) proposed by McDonald (1999) was used to provide evidence for the reliability of the measurements obtained after collecting the evidence for the construct validity of the measurement tool. In the present study, the use of the McDonald ω coefficient was preferred because it was aimed to obtain more consistent estimates in such measurements (Osburn, 2000) since the factor loads of the variables are different from each other. At the end of the analysis, the McDonald (ω) coefficient was found to be 0.972 (95% Confidence Interval: 0.963-0.978). As a result, evidence regarding the reliability of the measurements and the validity of the inferences based on the measurements were collected and the analysis of the hypotheses of the research was initiated.

Data Analysis

For data analysis, the many facet Rasch model was used. One-dimensionality, local independence, and model data fit are required for consistent and unbiased estimates of the

measurements obtained using the many facet Rasch analysis. Testing these assumptions serves the reliability and validity of the measurements. As stated in the section of collecting data for one-dimensionality as the first assumption, it was determined that the analytical graded scoring key has a single factor. In other words, it provides one-dimensionality. Since the one-dimensional measurement tool indicates local independence, it is accepted that local independence is also provided. Finally, standardized residual values were examined for model-data fit. It was stated that the number of standardized residual values outside the ± 2 range should not be more than 5% of the total number of observations, and the standardized residual values outside the ± 3 range should not be more than 1% of the total data number to ensure model data compliance (Linacre, 2017). When the standardized residual values were examined, it was found that there were 202 (4.89%) values in the ± 2 range and 17 (0.41%) values in the ± 3 range, and it was concluded that the model data fit was at an acceptable level (total number of observations 8x6x86 = 4128).

After the essays written by the students were scored by eight raters according to the analytical rubric, the average of the scores given by eight raters to each criterion for each student was calculated. Later, a two-stage clustering analysis was performed using these average scores and it was found that the participating students were divided into three groups according to their academic writing skills (Clustering quality = 0.710 and average Silhouette coefficient = 0.614). These groups were named as insufficient, moderate, and satisfactory according to their average scores.

Results

This study aims to investigate the effect of the differential rater function, which is one of the rater errors involved in the measurements as a result of the interactions between rater x student qualities in the process of evaluating the academic writing skills of middle school students. In this regard, both rater mismatches (such as severity and leniency) and the level of interference of the differential rater function in the measurements in the performance evaluation process were examined. The findings of the research are presented in subtitles in parallel with the sub-problems.

Findings with respect to the differential rater function at a group and individual level for rater x student efficiency interactions

The first sub-problem of the study was "Do the raters score the students according to their academic writing skills (unsuccessful, moderate, successful) and show the differential rater function at the group and individual level?". To find an answer to the question, rater x student competence interaction was made in the many facet Rasch model, and the results regarding the group and individual levels, respectively, are given in Table 3 and Table 4.

Table 3. Group-level DRF statistics for student competence x rater interaction

Interaction type	Explained variance (%)	Chi-square	df	p
Rater x student proficiency	0.11	23.70	24	0.48

According to Table 3**,** it was determined that the rater function, which differs statistically at the group level in the joint interactions between the rater and the student's proficiency level, was not shown by the raters ($p > 0.05$). The very low percentage of variance explained by the interaction effect supports the result reached.

Since the fact that the rater function differential at the group level does not appear does not

mean that it will not occur at the individual level, it is important to examine the statistical indicators at the individual level for the validity and reliability of the measurements. In this context, the statistical indicators of rater x student competence interactions at the individual level are given in Table 4 for those found to be significant.

Table 4. Meaningful Individual-level DRF statistics regarding the interaction between rater x and students' competency

Rater	Student qualification	Infit MSQ	Outfit MSQ	Observed Score	Expected Score	Bias Size	Model S.E.	t
R8	Moderate	1.00	1.00	275	288.57	-0.54	0.19	-2.78
R8	Unsuccessful	0.90	0.90	397	381.13	0.25	0.13	1.98

In Table 4, it was found that two of the three interactions of R8 numbered rater showed DRF. When the measurement report regarding the rater facet given in Annex 1 is examined, it is seen that the rater number R8 is the most generous rater. When the literature is reviewed, it is observed that the most generous or strict raters generally show DRF (Matsuno, 2009; Wolfe, & McVay, 2012). The graphical representation of rater x student competency interactions is shown in Figure 2.

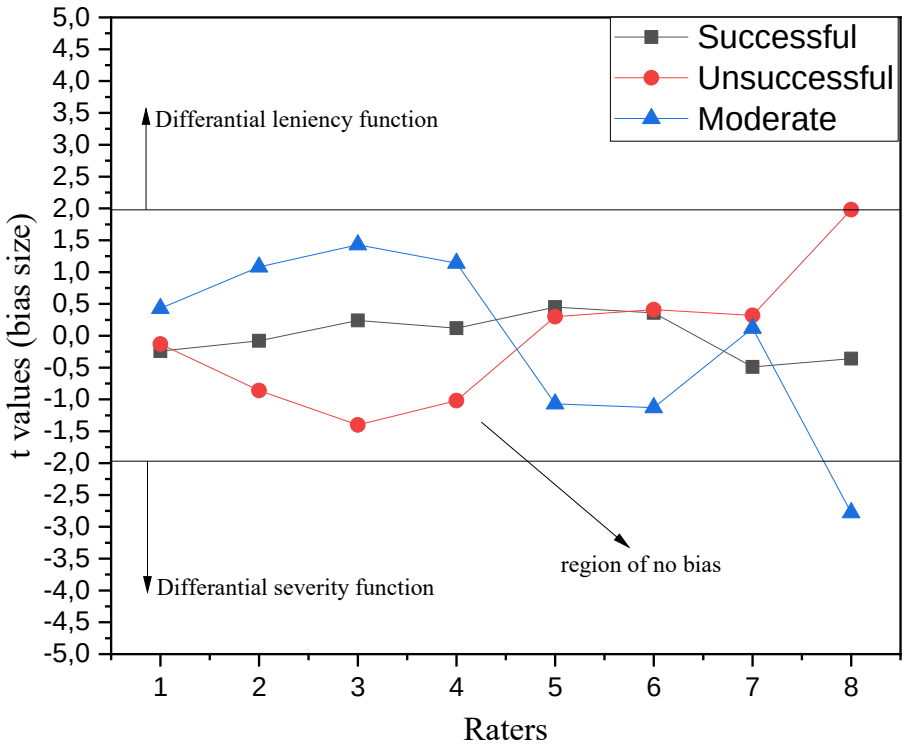

Figure 2. Rater x student efficacy interactions

When Figure 1 is examined, it is seen that successful students are exposed to bias less than other students. While rater number eight shows bias towards unsuccessful and moderate students, this is not the case for successful students. It is striking that especially moderate students are exposed to more bias.

After examining the DRF status of the raters according to the student's proficiency in academic writing skills, the effect of the student's gender on the rater's status of showing DRF in the evaluation process was examined. For this purpose, the second sub-problem of the study, "Do raters score the students according to their gender status (female, male) and show the different

rater function at the group and individual level?" was proposed in order to scrutinize the interactions between rater x student's gender. Group-level statistical indicators are shown in Table 5.

Table 5. Group-level DRF statistics of rater x student gender interaction

Interaction type	Explained Variance (%)	Chi-square	df	p
Rater x student's gender	0.02	7.10	16	0.97

According to Table 5, it was determined that the rater function, which differed statistically at the group level in the joint interactions between the rater and the student's gender, was not shown by the raters ($p > 0.05$). The very low percentage of variance explained by the interaction effect supports the result reached.

After examining the statistical indicators related to the group level of the rater x student's gender interaction, the statistical indicators at the individual level were examined and it was concluded that no interaction was significant, in other words, there was no DRF at the individual level. Graphical representation of rater x student gender interactions is given in Figure 3.

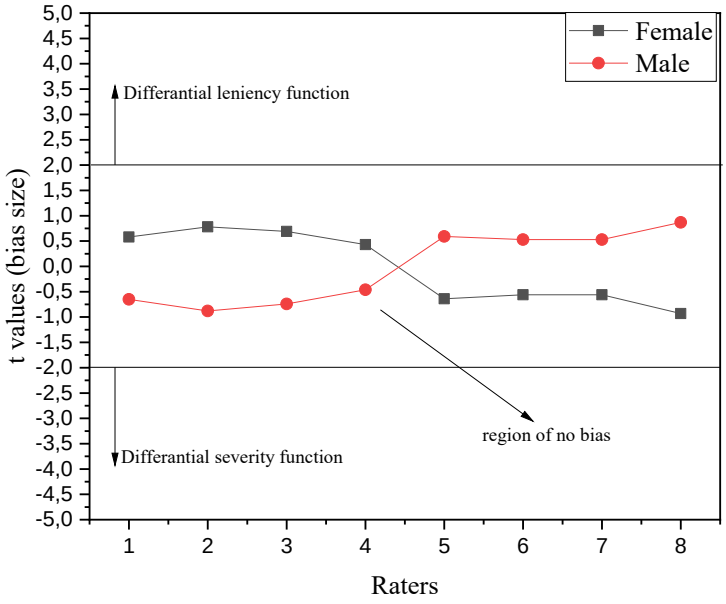

Figure 3. Rater x student gender interactions

Figure 3 shows that raters do not have rater behaviours that differ according to student gender. However, it is seen that there is a significant mismatch between the first four raters and the last four raters. The first four teachers' working in the state and the last four in the private school may be effective in this situation.

Discussion and Conclusion

In the study, it was aimed to reveal the interference level of the differential rater function, which is one of the rater errors in the process of evaluating the academic writing skills of students in primary education. previous literature suggests in relation to this that DRF is mostly investigated in the process of evaluating academic writing skills in the second language (Eckes, 2005; Engelhard, & Myford, 2003; Farrokhi, Esfandiari, & Schaefer, 2012; Johnson and Lim, 2009; Kondo-Brown, 2002). With the aim of determining the level of involvement of DRF in the measurements in the process of evaluating Turkish students' academic writing skills in their mother tongue, the current study hopes to contribute to the national literature and guide

future research. First, in the evaluation process of the raters, the DRF status of the students according to their academic writing skills were examined. Findings show that DRF at the individual level is involved in the measurements while DRF at the group level does not interfere with the measurement. Additionally, it was unearthed that the level of DRF involvement in the measurements of successful students was the lowest. In the study conducted by Wolfe and McVay (2012), it was found out that 40 raters who are especially generous or strict showed DRF during the evaluation process of 120 students' compositions. In the same vein, the present study revealed that the raters showing DRF were also the most generous ones. However, in the study conducted by Kondo-Brown (2002) it was found that the DRF levels in the measurements of successful and unsuccessful students were the highest. The main reason for this situation stems from the assessment of students' academic writing skills in the second language in the study conducted by Kondo-Brown (2002), while the current study evaluates students' academic writing skills in their mother tongue. In the process of evaluating students' academic writing skills, it is seen that DRF's involvement in measurements according to student competencies is an expected result and the literature supports this result. Another important result of this study is that teachers working in private and public schools exhibit different scoring behaviours. It is thought that this situation may be due to the tendency of private school teachers to give students higher grades. Researchers state that teachers tend to give higher grades to students studying at private schools. Among the main reasons for this situation are private school owners' who wish to persuade the students to stay in their schools by putting grade-pressure to the teachers in order to meet the high grade demands of the parents. Also they wish to give the image as if it is a succesful school by keeping the grade point average of the school high (Garipağaoğlu, 2015; Gürler 2020). Berberoğlu and Kalender (2005) state that students studying in private high schools get higher scores in university entrance exams with their high school grades than students studying in public high schools.

In the evaluation of the academic writing skills of the students in their mother tongue, the DRF status of the raters according to the gender of the students was examined and it was found that there was no significant interaction. In other words, the gender of the students was not effective in the performance evaluation process. In the study conducted by Engelhard and Myford (2003), DRF involvement in the measurements according to the gender of the students in the performance evaluation process was examined and it was found that there was no significant interaction. In the study conducted by Gyagenda and Engelhard (2009), it was found that although female students have better academic writing skills due to their nature, rater x student gender interactions are not significant. Thus, the results of this study seem to be consistent with the previous literature in this area.

In this study, the fact that rater x student gender does not give meaningful interactions serves the validity of the measurements. However, when the interactions related to rater x student gender were examined, it was found that teachers in private and public schools again exhibit different behaviours. When these findings obtained within the scope of the study are considered together, it is important to consider this situation in cases where the scoring behaviours of teachers in private and public schools differ and teachers working in both school types are raters.

To conclude, this study revealed that DRF interfered with the measurements in rater x student competence interactions, yet DRF did not interfere with the measurements in rater x student gender interactions. In other words, while teachers or raters tended to score differently according to their writing skills in the process of evaluating students' academic writing skills in their mother tongue, there was no difference regarding their gender.

Appendix 1

Table 1. Rater Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	N	Rater
1504	516	2.91	3.01	.98	.09	1.02	.2	1.06	.5	.92	.59	55.9	50.6	8	R08
1497	516	2.90	3.00	.92	.09	.91	-1.4	.95	-.5	1.00	.59	57.6	50.9	5	R05
1483	516	2.87	2.98	.82	.09	.88	-1.9	.90	-.9	1.07	.60	58.6	51.4	6	R06
1466	516	2.84	2.96	.69	.09	.90	-1.7	.92	-.9	1.09	.59	59.2	52.0	7	R07
1292	516	2.50	2.68	-.60	.09	1.09	1.4	1.06	.8	.94	.60	61.8	52.1	4	R04
1273	516	2.47	2.63	-.75	.09	1.01	.1	.94	-.8	1.03	.62	62.4	51.5	3	R03
1238	516	2.40	2.55	-1.02	.09	1.08	1.3	1.07	.9	.92	.61	59.4	50.1	1	R01
1235	516	2.39	2.55	-1.04	.09	.99	.0	.91	-1.1	1.04	.62	59.7	50.0	2	R02
1373.5	516.0	2.66	2.79	.00	.09	.99	-.3	.98	-.3		.60			Mean (Count: 8)	
115.7	.0	.22	.20	.87	.00	.07	1.2	.07	.8		.01			S.D. (Population)	
123.7	.0	.24	.21	.93	.00	.08	1.3	.07	.9		.01			S.D. (Sample)	

Model, Populn: RMSE .09 Adj (True) S.D. .86 Separation 9.83 Strata 13.44 Reliability (not inter-rater) .99
Model, Sample: RMSE .09 Adj (True) S.D. .92 Separation 10.51 Strata 14.35 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 777.4 d.f.: 7 significance (probability): .00
Model, Random (normal) chi-square: 6.9 d.f.: 6 significance (probability): .33
Inter-Rater agreement opportunities: 14448 Exact agreements: 8572 = 59.3% Expected: 7379.4 = 51.1%

References

Aslanoğlu, A. E, & Kutlu, Ö. (2003). Research on rubric in evaluating the presentation skills in education. *Ankara University Faculty of Educational Sciences Journal*, *36*(1-2), 25-36.

Berberoğlu G., & Kalender İ. (2005). Investigation of student achievement across year, school types and regions: The SSE and PISA analyses. *Educational Sciences and Practice Journal*, 4(7), 21-35.

Bond, T. G., & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Collins, J. L. (2000). Review of key concepts in strategic reading and writing instruction. J. L. Collins (Ed.), in *Cheektowaga-sloan handbook of practical reading and writing strategies* (pp. 5-10). Retrieved from http://gse.buffalo.edu/org/writingstrategies/PDFFiles/CHEEKTOWAGA-SLOAN.PDF

Du, Y.,Wright, B. D., & Brown, W. L. (1996, April). *Differential facet functioning detection in direct writing assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Eckes, T. (2005). Examining rater effects in test of writing and speaking performance assessments: A many-facet rasch analysis. *Language Assessment Quarterly*, *2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

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. (2019). Many-Facet Rasch measurement: Implications for rater-mediated language assesment. *In Quantitative Data Analysis for Language Assessment* (1st ed.) (pp.153-175). UK: Routledge.

Englert, C. S., & Mariage, T. (2003). The sociocultural model in special education interventions: Apprenticing students in higher-order thinking. In L. H. Swanson, K. Harris, & S. Graham (Eds.), *Handbook of Learning Disabilities* (pp. 450-467). New York: Guilford.

Erhardt, R. P., & Meade, V. (2005). Improving handwriting without teaching handwriting: The consultative clinical reasoning process. *Australian Occupational Therapy Journal*, *52*(3), 199-210. https://doi.org/10.1111/j.1440-1630.2005.00505.x

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93- 112.

Engelhard, 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* (1), i-60.

Engelhard, G., Jr. (2007). Differential rater functioning. *Rasch Measurement Transactions*, *21*, 1124-1125.

Englert, C. S., Raphael, T. E., Anderson Helene M., Anthony, L. M., & Stevens, D. D. (1991). Making strategies and self-talk visible: Writing instruction in regular and special education classrooms. *American Educational Research Journal*, *28*(2), 337–372. https://doi.org/10.3102/00028312028002337

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*, 70-77.

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

Garipağaoğlu, B.Ç. (2015). Private school sector and unethical achievement engineering practices. *Ahi Evran University Kırşehir Faculty of Educational Sciences Journal*, 16(3), 181-200

Goodrich, H. (1997). Understanding rubrics. *Educational Leadership*, *54*(4), 14-17.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, *5*(1), 13-34.

Gürler, M. (2020). Differences between public school and private school. *Kapadokya Education Journal*, 1(1), 1-6

Gyagenda, I. S., & Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement*, *10*(3), 225-246.

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English majör graduates. *Chinese Journal of Applied Linguistics*, *33*(2), 87-102.

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

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

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505. https://doi.org/10.1177/0265532209340186

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31. https://doi.org/10.1191/0265532202lt218oa

Kuan-Yu Jin & Wen-Chung Wang (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

Lam, S. S. T., Au, R. K. C., Leung, H. W. H. & Li Tsang, C. W. P. (2011). Chinese handwriting performance of primary school children with dyslexia. *Research in Developmental Disabilities*, 32, 1745-1756. https://doi.org/10.1016/j.ridd.2011.03.001

Linacre, J.M. (2018). A user's guide to FACETS Rasch-model computer programs. *Program manual 3.81. 0.* Chicago: MESA Press.

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives. Experts in assesment*. Thousand Oaks, CA: Corwin Press, Inc.

McDonald, M. B. (1999). Seed Deterioration: Physiology, Repair and Assessment. *Seed Science and Technology*, *27*(1), 177-237. Retrieved from https://ci.nii.ac.jp/naid/10025267238/

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

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*(3), 343-355. https://doi.org/10.1037/1082-989X.5.3.343

Şata, M. (2019). *Performans degerlendirme surecinde puanlayici egitiminin puanlayici davranislari uzerindeki etkisinin incelenmesi [The investigation of the effect of rater training on the rater behaviors in the performance assessment process].* Unpublished doctoral dissertation. Gazi University, Ankara.

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

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303-325. https://doi.org/10.1191/026553201680188988

Tamanini, K. B. (2008). *Evaluating differential rater functioning in performance ratings: Using a goal-based approach*. Unpublished doctoral dissertation. Ohio University, Ohio.

Wesolowski, B. C., Wind, S. A., & Engelhard, 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

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