European Journal of Foreign Language Teaching
ISSN: 2537 - 1754
ISSN-L: 2537 - 1754
Available on-line at: www.oapub.org/edu

doi: 10.5281/zenodo.1947027 Volume 3 | Issue 4 | 2018

DEFINING SEVERE GRADERS THROUGH MANY FACETED RASCH MEASUREMENT

Murat Polat[i]
Anadolu University,
Eskisehir, Turkey

Abstract:
Scoring language learners' writing exams is a difficult task for graders since many task-relevant or irrelevant variables such as the user-friendliness of the rubric, difficulty of the task, students' handwriting or grader characteristics (being too lenient or harsh) are involved in the process. To be able to gain valid and reliable scores, studying the variables that affect scoring procedures and seeking ways to control and minimize them are crucial concerns for institutions in order to assure their learners that their assigned scores are genuine and given in the least subjective way that could be possible. That is why analysing grader attitudes while scoring and identifying the stringent and lenient graders in the rater-pool is important not only to be able to set the best matches of graders where multiple scorings or cross-marking sessions are applied but for making those raters be aware of their scoring habits. In this exploratory study, 6 writing graders who had more than 10-year-expertise in grading writing voluntarily scored 20 student essays including two separate tasks. MFRM (Many Faceted Rasch Measurement) was used to explore graders' marking behaviours and discover how those behaviours affect test scores of language learners. Finally, results of the study showed that graders, while they all used the same rubric and had enough expertise in grading, have significant score differences and a significant level of stringency in scoring essays.

Keywords: testing, subjectivity, reliability, severe graders, lenient graders, Rasch analysis, rater effect

1. Introduction

In most educational settings, it is a well-known fact that the scores assigned to students' papers or performances unfortunately do not depend merely on students' test performance or success but on may other test-relevant or irrelevant factors. Among the factors are test difficulty, grader behaviour while scoring (being lenient or stringent), graders' attitudes to the scoring rubric or the extent to how much or how effective the

[i] Correspondence: email mpolat@anadolu.edu.tr

rubric is used, the time spent for marking, the purpose of marking, the impression of a student (students' identities, gender, cultural backgrounds or even the students' handwriting or examples they use while writing), their physical qualities (in cases where students are visible) etc. Therefore, such variables and surely grader behaviours (possibly the most important and the popular issue in research) must certainly be taken into consideration to be able to assess students' true test performances in a valid and reliable way since those graders can vary in their testing and assessment interpretation, rubric use and level of harshness while grading (Coniam & Falvey, 2007; Lane & Stone, 2006). It is an undeniable fact all these differences, which stem from human factor could well contribute to a number of measurement errors, to unreliable and invalid testing and the lack of fairness in the valuation of students' skills which may affect directly or indirectly manipulate many educational decisions. Thus, it is crucial to study and identify grader behaviours to be able to make better testing practices and decisions; therefore, the aim of this study is to analyse grader differences in terms of severity and leniency they displayed while assessing students' writing abilities using Many Faceted Rasch Measurement Model which is a useful tool in determining grader characteristics through many facets. (Linacre, 1989)

2. Literature Review

In the assessment of foreign language skills, implementation of the writing tests is a commonly used powerful tool that could reflect a vivid picture of the test taker's knowledge and language skills in a pre-determined context or domain which is tested in the target language. Although it is mostly time consuming, labor intensive and costly, the use of writing texts is indispensable for most institutions since they provide a good sum of evidence for the productive qualities of the language learners. Anastasi (1988) states that writing tests are surely necessary, however, the grading process in these written tests is not an easy task because of some variables like the psychology of the test taker, task difficulty, rater behavior and the quality of the scoring rubric. When all these factors are taken into account, it is no surprise that assessment of writing is highly likely to result in a number of errors not least because of the human factor involved at each step in the grading process. Wu and Tan (2016) warn that language learners' test scores should reflect their true language skills as precisely as possible since those scores could affect important decisions that could affect a person's future significantly in most cases.

The implementation of training or norming sessions can be an effective tool to optimize graders' consistent use of scoring rubrics (either analytic or holistic); however, as McNamara (1996) noted graders have personal perceptions of scoring behaviors which are not easily changed. A number of other studies (Lane & Sone, 2006; Lumley & McNamara, 1995; Wolfe, 2004) revealed the same fact, even implementing proper training sessions or providing sound rubrics do not completely resolve the rater differences in grading, they just reduce those errors or differences to some extent but they still exist. Thus, it is rather a utopia to guarantee a grader-error free writing exam

scoring or a subjective assessment process which is completely free of task relevant or irrelevant factors. The most practical and pragmatic action to take here is to make the graders aware of their grading behaviors and let them change whatever is necessary to make their grading fair and true considering the descriptors of the assigned tasks and given rubrics.

Think of a language school where hundreds even thousands of student papers are scored by writing graders, do you think that they all keep the same line and score the papers objectively? Some raters are, by their nature, too lenient while scoring ,which might be a positive attitude from the students' perspective; however, this might lead to wrong decisions about the students' language skills, it might cause problems especially if it is a placement test or if the language program is an intensive one that makes decisions on learners' abilities by those tests. Myford and Wolfe (2004) state that lenient graders are those who mostly assign on average higher marks than predicted results compared to the possible scores given by other raters, and it is a measurement error even if the students are happy with the results. On the other hand, Congdon and McQueen (2000) asserted that rater stringency or severity is the relative likelihood of graders to assign lower grades which is indeed a phenomena in testing which can turn scoring writing papers into heads or tails, if the rater is harsh you lose, you win if it is a lenient one regardless of the true quality of your performance. Considering the fact that not many language schools have the essential number of raters to cross-mark the students' papers, necessary precautions must be taken to ensure no student is over/disadvantaged by the random allocation of its work to a lenient or stringent grader, no matter how experienced, educated or well-trained that rater might be.

Messick (1995) underlined the rater error in scoring writing and stated that biased exam scores result in construct irrelevant traces of assessment errors which reduce test reliability and validity. That is why rater effect is highly important not only in terms of fairness of the assessment but also in terms of its validity and reliability. Also, Lumley (2005) and Eckes (2005) studied rater effects in foreign language writing assessment, and they both revealed that rater severity or leniency caused significant score differentiation. In sum, considering all these studies, it is evident that human effect is an inevitable part of assessment which should be managed rather than trying hard to eliminate it completely. It could be, thus, wiser to seek ways to identify these lenient or severe graders and for us as educators, the ethical obligation dictates that necessary precautions must be taken to control and compensate the scoring effects in grading students' performances.

MacMillan (2000) suggested the use of *Rasch Models* to study the effect of lenient or stringent graders on students' scores. Likewise, rather than the use of *Classical Test Theory (CTT)* in detecting inter-rater variability and rater effect in grading, other researchers (Kondo & Brown, 2002; Lunz et al., 1990; Park, 2004; Prieto, 2011; Razak et al., 2012; Tyndall & Kenyon, 1996) propose the use of *Generalizability Theory (GT)* and the *Multi-Faceted Rasch Measurement (MFRM)*. MFRM is more advantageous since it enables the researcher to analyze the scoring behaviors of various raters on different tasks (Boone, 2016; Linacre, 1989) and thus permits the researcher see if the scoring

components in rubrics need to be revised or changed to obtain reliable and valid results. MFRM is also used to obtain true measures from raw marks on a number of variables affecting the scoring quality of a writing test. Di Nisio (2010) stated that the MFRM model is a successful extension of the Rasch models and can be very useful to examine rater effects when scoring foreign language writing or speaking exams.

All in all, the aim of this study is to analyze rater differences while scoring writing papers of English learning students in terms of stringency or leniency of the raters by using the MFRM model. By using the findings gathered from this study, it is planned to identify rating behaviors of the raters and considering the components of the rubric and the variety of the scores. If a significant scoring difference is observed, it will also be discussed which precautions could be taken to manage the rater effects and minimize them for gaining more reliable and valid scores in testing writing.

3. Method

This exploratory study aims to investigate raters' scoring differences while assessing foreign language writing tasks in terms of rater leniency/stringency. In this study, it is intended to answer the following research questions using the data obtained from the scorings:

1. Do the raters' marks differ significantly although they use the same scoring guidelines?
2. Do raters significantly differ from the others in terms of leniency/stringency?
3. Do the scoring components differ from the others in terms of difficulty?

3.1 Participants

There were two groups of participants in this study. The rater group consisted of six English language instructors who were working at a language school of a state university in Turkey. All but one had MA degrees in English Language Teaching and were marking students' papers for more than 10 years. Those raters are the members of the rater pool of the language school whose inter and intra-rater reliability levels are supposed to be satisfactory. As for the students, 22 intermediate level language learners (aged between 18-21) participated to the study; however, two of them were excluded from the study since they did not obey the writing rules of the school in terms of minimum word limit and writing on the relevant task. Finally, a total of 20 university students and 6 raters joined voluntarily participated to this study.

3.2 Instruments

A writing exam which had two separate tasks was used in this study. In the first task, the students were asked to write a *Process Paragraph* (a brief summary in 100-120 words to describe their preparation process to the exam) about what they did to prepare for the university exam. The second part's question was "Do you think that building a nuclear power plant in Turkey is a good idea?"

In the *Opinion Paragraph,* students were supposed to express their opinions and provide necessary reasons and examples related to the task. The scorings were done by an analytic scoring rubric which has five components (content, organization, grammar, vocabulary and mechanics). The rubric was developed in the same language school, and the graders are familiar with this rubric since they all used it for assessment of writing purposes many times. Each component in the rubric has a score range from 4-0 which means 4 for the excellence, 3 for good performance, 2 for average, 1 for weak and 0 for the poor quality in the related component.

3.3 Procedure

All the participants contributed to the study voluntarily. First, the students wrote their papers in a 75-minute writing session. After all the written works of the students were collected, their names and personal info on the papers were hidden, each paper was numbered from 1-22 (two papers were excluded later and 20 papers were used for the study) and all the papers were photocopied to provide a copy for each to make all the raters score the same samples. The scoring session took 4 hours and all the graders scored the papers individually. The score data gathered from the participants were computed and analyzed using the FACETS (Linacre, 2009) program which enables its user to run MFRM analysis to study parameter estimation, necessary sampling for conjoint measurement, analysis of infit and outfit degrees to obtain fit indices of the distribution properly.

4. Results and Discussion

In the analysis of the data gathered from the raters, MFRM model is used since it is an advanced sub-model of the Partial Credit Model for polytomous items in which a student's ability is scored by using a criterion or criteria, each of which is composed of a set of related and pre-arranged categories (Linacre & Wright, 2002). Prieto and Nieto (2014) suggest that this model can be applied to educational assessment cases in which there could be seen a number of dependent or independent variables such as student performance, task difficulty, rater difference or scoring rubrics which can ultimately lead to measurement error. It is thought that this model could be used in the detection of measurement error and finding out the role of each facet to the logit or logarithm of the ratio between the possibility that a performance to be assessed will receive one score on the rubric (let's say, 4) and the possibility of that same performance receiving another score which might be lower (let's say, 3).

The four faceted Rasch model presented by Prieto and Nieto (2014, p: 387) is,

$$log\ (Pnijlk\ /Pnijl(k\text{-}1) = Bn - Rj - Di - Fjk$$

where Pnijlk is the possibility of task n being scored k by grader j; Pnijl(k – 1) is the possibility of task n being rated k – 1 by grader j; Bn is the skill of the student as shown in the quality of the task; Rj is the stringency of the grader; Di is the difficulty of

the test and Fjk is the difficulty of the scoring rubric step comparative to the previous step.

Linacre (2003) recommends that to be able to run a basic analysis with MFRM, less than 5% of the standardized values (z-scores) in the data set should be equal to or more than 2, or less than 1% of the standardized values (z-scores) should be equal to or more than 3. It was estimated that out of 1800 total data in this study 34 (0.02%) were equal or more than 3 and out of 1800 total data 55 (0.03 %) were equal or more than 2 and those results displayed that the model was fit for the analysis since Linacre (2003) defends the idea that not only too much, but also too little, observed "error" variance might threaten the validity of the assessment.

The variable map in Figure 1 displays an overall view of the total data gathered from this study including the measurement units (column 1) between 5 and -4 logits, paper quality (column 2), grader stringency (column 3), task difficulty (column 4), component difficulty (column 5) and the functionality of the scoring components (column 6-11) respectively. Each asterisk in column two stands for a single student paper and seeing the wide distribution, the difference of their assigned scores could be predicted although they were all intermediate level language learners.

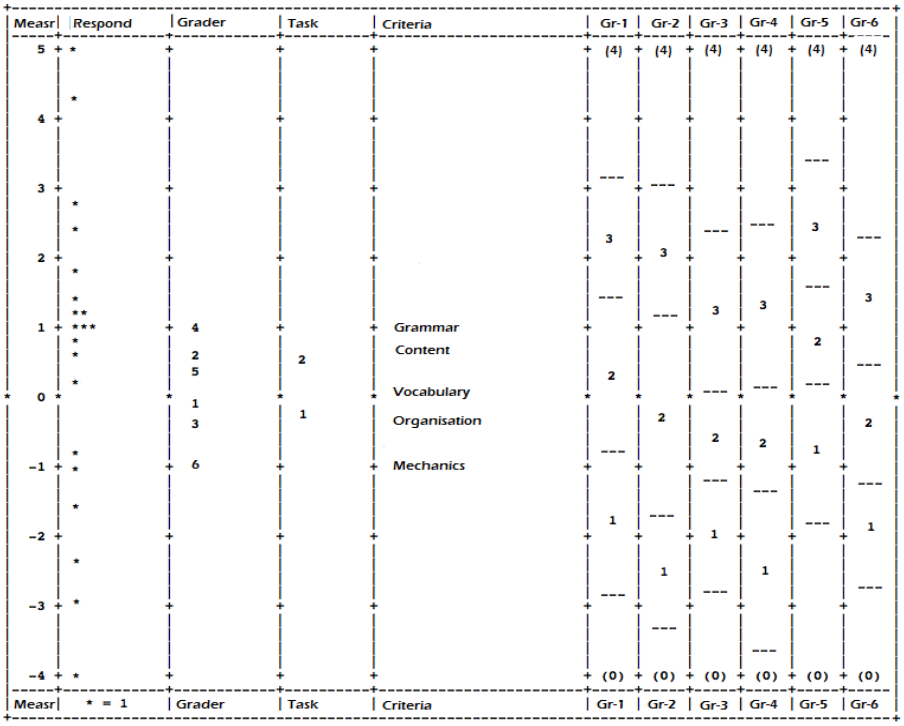

Figure 1: The variable map presenting the rank of papers, graders, tasks and components

The grader column (column 3) in Figure 1 displays the rank of graders in terms of leniency/stringency. The map reveals that the severe to lenient grading range is from around +1 to -1 logits and Grader 4 was found to be the most lenient whereas Grader 6 seemed to be the most stringent rater. The fourth column shows that the second task in which students were asked to discuss the advantages/disadvantages of building nuclear power plants was found to be more difficult than the first task in which they were asked to write how they prepared for the university exam. As for the difficulty of the

components which was presented in the criteria column (column 5), grammar was found to be the most difficult component for students since writing accurate and grammatically correct English sentences is highly appreciated in the language school whereas another important concern mechanics (capitalization, punctuation, spelling etc.) was identified to be the easiest component in the assessment of writing. The detailed analysis of each variable will be given in the following tables.

Table 2: Student papers' measurement report (MFRM)

Obsvd Score	Obsvd Count	Obsvd Average	Fair Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Paper
353	90	3.8	3.87	1.26	.15	1.1	1	1.0	1	20
345	90	3.7	3.78	1.10	.14	1.3	1	1.2	1	2
336	90	3.6	3.71	.98	.14	1.0	0	0.9	0	5
335	90	3.6	3.70	.97	.14	0.9	-1	1.0	-1	7
330	90	3.6	3.68	.88	.14	1.2	1	1.2	0	19
328	90	3.5	3.64	.86	.14	1.1	1	1.0	1	8
327	90	3.5	3.62	.85	.14	1.1	1	1.1	0	14
325	90	3.5	3.59	.77	.13	1.0	0	1.0	0	18
320	90	3.5	3.59	.73	.13	1.2	1	1.1	1	16
310	90	3.4	3.52	.61	.13	1.3	1	1.2	1	15
306	90	3.3	3.46	.52	.13	1.0	1	1.0	0	1
294	90	3.2	3.41	.35	.12	0.9	0	0.9	0	10
293	90	3.1	3.38	.34	.12	0.9	0	0.8	0	9
292	90	3.1	3.35	.31	.12	0.8	-1	0.7	-1	12
291	90	3.1	3.31	.30	.12	1.0	0	1.0	0	6
289	90	3.1	3.28	.29	.12	1.1	1	1.0	1	17
285	90	3.1	3.23	.27	.12	0.7	-2	0.7	-1	13
281	90	3.0	3.20	.24	.12	1.2	1	1.3	1	11
264	90	2.8	3.02	.05	.12	1.3	1	1.2	1	3
256	90	2.8	3.01	-.01	.11	1.0	0	1.1	0	4
308.0	90.0	3.3	3.28	.60	.13	1.2	0.2	1.0	-0.0	Count:20
27.0	0.0	0.3	0.26	.38	.01	0.2	1.1	0.2	1.0	S.D.

RMSE (Model) .11 Adj S.D. .32 Separation 2.78 Reliability .88
Fixed (all same) chi-square: 178.4 d.f.: 19 significance: .00
Random (normal) chi-square: 18.3 d.f.:18 significance: .38

The detailed analysis of the students' written works was presented in Table 1. Students were asked to write two different paragraphs and the mean scores of the two paragraphs which were assigned by the raters were examined. The 20th participant's written work, paper 20 was found to be the most successful whereas paper 3 had the weakest performance out of those 20 papers. The RMSE (Root Mean Square Standard Error) value shows the standard error mean value for the whole data except the outliers and this value was found as 0.12, which means that standard error mean was remarkably low in this analysis. In order to justify the RMSE value, the Adjusted Standard Deviation was also checked and it was found as 0.33 which is well below the critical level 1.0 (Wright & Linacre,1994).

The reliability measurement in Rasch analysis is the same with the measure used in the techniques such as KR 20-21 or Cronbach Alpha tests; it is a measure between 0-1, and the higher the better. Therefore, reliability is the portion of the overall variance in a measure which is true score variance; in other words, a test's reliability is defined as the ratio of true score variance to observed-score variance (Wright & Masters, 1982). The reliability of the analysis given in Table 1 was found as 0.88, and this statistical analysis could be accepted as highly reliable since it is more than 0.85. As for the quality of the students' works, the hypothesis "students' written works have no statistical difference in quality" was rejected due to significant quality differences among students' works ($\chi 2$ =178.4, df = 19, $p<0.05$).

Another important advantage of using MFRM is that it gives statistical infit and outfit values of the various facets. "*Infit*" means inlier-sensitive or information-weighted fit and this is more delicate to the pattern of replies to items targeted on the person or others whereas *"Outfit"* means outlier-sensitive fit which is more delicate to replies to items with difficulty far from a person, and others .Wright and Linacre (1994) reported the critical limits as the values between 0.6 - 1.4. Considering those limits, none of the infit or outfit values in Table 1 was exceeding the given limits which reveals that the scores assigned to the students' written works were fit to the model.

Table 2: Grader measurement report (MFRM)

Obsvd Score	Obsvd Count	Obsvd Average	Fair Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Grader
1131	300	3.8	3.86	.45	.08	1.1	1	1.0	1	4
1098	300	3.6	3.72	.24	.07	1.0	0	1.0	0	2
1087	300	3.6	3.70	.20	.07	1.7	4	1.6	4	5
1086	300	3.5	3.61	-.21	.06	0.9	-1	1.0	-1	1
918	300	3.0	3.16	-.39	.06	0.8	-2	0.7	-1	3
816	300	2.7	2.84	-.68	.05	1.1	1	1.0	1	6
1022.7	300.0	3.4	3.48	.00	.07	1.0	0.2	1.0	-0.3	Count:20
110.1	0.0	0.3	0.33	.21	.01	0.2	2.1	0.2	2.5	S.D.

RMSE (Model) .08 Adj S.D. .40 Separation 6.28 Reliability .91
Fixed (all same) chi-square: 246.4 d.f.: 5 significance: .00
Random (normal) chi-square: 5.3 d.f.:4 significance: .28

The detailed analysis of the six graders who scored all the papers in this study was presented in Table 2. It should be reminded that all the graders participated in the study had at least ten years of grading experience and were highly qualified in assessment. The results showed that Grader 4 was the most lenient grader who had assigned 3.801 points on average to each of the components in the rubric out of 4 points. Grader 6 was the most stringent grader who had assigned 2.698 points on average to each of the components in the rubric out of 4 points. The RMSE (Root Mean Square Standard Error) value shows the standard error mean value for the whole data except the outliers and this value was found as 0.8 which means that standard error mean was remarkably low in this analysis. In order to justify the RMSE value, the Adjusted Standard Deviation was also checked, and it was found as 0.40 which is below the critical level 1.0. The

reliability of the analysis given in Table 2 was found as 0.91, and this statistical analysis could be accepted as highly reliable since it is more than 0.85 (Wright & Linacre,1994). As for the scoring behaviors of the participant raters, the hypothesis "raters have no statistical difference in their scoring behaviors" was rejected due to significant scoring differences among the raters who contributed to the study ($\chi 2$ =264.4, df = 5, $p<0.05$). When the infit and outfit values for the raters' scoring performances were observed, it could be said that all the graders but Grader 5 are within the pre-defined limits and can score the papers within a high inter-rater reliability range; however, Grader 5's both infit (1.7) and outfit (1.6) values are over the critical limit (1.4) and this leads us to the result that the scoring behavior of the rater is significantly different from the other five raters and Grader 5's scores in this data set are not reliable judgements. This finding is important considering the fact that under normal conditions raters feel that their judgements are fair and in parallel with the descriptors written in scoring rubrics; however, it is a fact that exercising various scoring practices could not only make the raters fit for scoring but also would give the administrators a chance to see which raters may need more norming or training to make better decisions.

Table 3: Rubric components' measurement report (MFRM)

Obsvd Score	Obsvd Count	Obsvd Average	Fair Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Componet
286	120	2.2	2.36	1.06	.07	1.0	1	1.0	1	Grammar
356	120	2.8	2.84	.43	.09	1.3	0	1.2	0	Content
387	120	3.0	3.14	.08	.11	0.9	1	0.9	1	Vocabulary
431	120	3.2	3.31	-.37	.12	1.2	-1	1.1	0	Organisation
488	120	3.6	3.62	-.96	.13	1.0	-1	0.9	-1	Mechanics
389.6	300.0	3.4	3.05	.00	.09	1.1	0.1	1.0	-0.1	Count:20
54.7	0.0	0.3	0.38	.48	.02	0.3	2.3	0.3	2.4	S.D.

RMSE (Model) .10 Adj S.D. .56 Separation 5.48 Reliability .93
Fixed (all same) chi-square: 346.2 d.f.: 4 significance: .00
Random (normal) chi-square: 15.2 d.f.:3 significance: .32

The final analysis given in Table 3 is related to the components of the scoring rubric which was used by the raters in the study. It should be reminded that this rubric was developed by the testing unit of the language school where all the raters work and grade students' papers. The reliability analysis of the rubric was made before by the testing unit and it was reported as a reliable tool in assessment whose reliability was computed as 0.80. This fact is particularly important since making a component analysis of a scoring guide whose reliability is unknown or below 0.6 would be a serious mistake. The results in Table 3 showed that "grammar" was the most difficult component by which raters assigned 2.198 points on average out of 4 points. As it was mentioned before, intensive grammar teaching and the expectancy of students' using correct grammar and accurate forms are common in many countries where English is taught as a foreign language like Turkey. In such settings, unfortunately most of the language tests measure mainly the grammar skills of students; therefore, whether it is a writing or speaking test, producing grammatically correct sentences is highly

appreciated; that is why in a Turkish context the priority of the accurate language use and grammar is no surprise. On the other hand, including many important conventions of writing such as capitalization, punctuation or spelling "mechanics" component of the rubric was found to be the easiest one. Raters assigned 3.632 points on average out of 4 points which is really high when compared to the grammar component. The RMSE (Root Mean Square Standard Error) value shows the standard error mean value for the whole data except the outliers, and this value was found as 0.10 which means that standard error mean was remarkably low in this analysis. In order to justify the RMSE value, the Adjusted Standard Deviation was also checked, and it was found as 0.56 which is below the critical level 1.0. The reliability of the analysis given in Table 3 was found as 0.93 and this statistical analysis could be accepted as highly reliable since it is more than 0.85 (Wright & Linacre, 1994). As for the different components of the rubric which have equal score weights, the hypothesis "components have no difference in terms of difficulty" was rejected since there appeared significant mean score differences among the five components ($\chi 2$ =346.2.4, df = 4, $p<0.05$). When the infit and outfit values of the components were taken into account, it could be said that none of the infit or outfit values of the components in Table 3 was exceeding the accepted limits (0.6-1.4) which reveals that the scores assigned to the students' written works were fit to the model.

5. Recommendations

This study aimed to reveal scoring differences of expert graders and their potential leniency or stringency in grading although they all use the same scoring rubric and they have all worked as part of a team for many years. It was a voluntary based study; that's why only 6 raters contributed to it and it is a well-known fact that in such statistical studies the number of the participants either the raters or the rates is highly important, the more the better. Thus, a replication of this study with more participants could be recommended. Another recommendation is for the testing units; the use of MFRM might be very useful in defining rater behaviors and could give more insights in true scoring of the students' language skills. The last but not the least, the aim of scoring is another important concern. If a similar study is designed under actual conditions where raters score not for the use of a researcher to analyze in an empirical study but for assessing students' writing skills under exam conditions, it might give more authentic and case sensitive results for the administrations and examiners.

6. Conclusion

In this exploratory study, raters' scoring behaviors and their levels of leniency and stringency in scoring were examined. The results obtained from MFRM (Multi-Faceted Rasch Measurement) analysis revealed that there were significant quality differences among students' written works although they were all intermediate level language learners ($\chi 2$ =178.4, df = 19, $p<0.05$). Another finding related to the raters' scoring

behaviors of the participants was that their mean scores had significant differences ($\chi 2$ =264.4, df = 5, $p<0.05$) and out of six expert graders, Grader 5 scored the 20 papers differently from the other raters and could not be considered to have reliable scoring judgements in this scoring process. Finally, out of five different components (content, organization, grammar, vocabulary and mechanics) of the scoring rubric that the raters used while rating, "grammar" component was found to be the most difficult (2.198 points on average out of 4 points) by which graders were most stringent while scoring, "mechanics" component which could have been a component that raters were more critical was found to be the easiest one (3.632 points on average out of 4 points) by which graders were most lenient while scoring. The reality that language teachers still value correct grammar use and accuracy in writing more than the other qualities like task achievement and organization in writing is an important finding driven from data. This could lead to education programmers make an important discussion on what should be favored most in the assessment of writing in a foreign language program, content or the form. Another important finding was that no matter how experienced or well-trained the raters might be, there are harsh or lenient graders in rater pools of language schools, and those raters should be examined and be identified periodically by the testing units and should be informed that they score differently compared to their colleagues, which indeed may cause unfair exam results and thus, may have serious effects on their students' academic lives.

References

Anastasi A, 1988. Psychological Testing, 6th Edition. New York: Macmillan.

Boone, W.J., 2016. Rasch Analysis for instrument development: Why, When and How?. CBE Life Science Education. 15-1/7, 2016.

Congdon, P.J., McQueen, J. 2000. The stability of rater severity in large-scale assessment programs. Journal of Educational Measurement, 37(2), 163–178.

Coniam, D., Falvey, P. 2007. High-stakes testing and assessment. In J. Cummins & C. Davison (Eds.), International handbook of English language teaching (pp. 457–471). New York: Springer.

Di Nisio, R. 2010. Measure school learning through Rasch Analysis: the interpretation of results. Procedia - Social and Behavioral Sciences, Volume 9, 2010, Pages 373-377. https://doi.org/10.1016/j.sbspro.2010.12.167

Eckes, T. 2005. Examining rater effects in TestDaF writing and speaking performance assessments: A multi-faceted Rasch analysis. Language Ass. Quarterly, 2(3), 197–221.

Kondo-Brown, K. 2002. An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. Lang. Test, 19, 1-29. http://dx.doi.org/10.1191/02655322o2it218oa

Lane, S., Stone, C.A. 2006. Performance Assessment. In R. L. Brennan (Ed.): Educational Measurement (pp 387-431). Wesport, CT: ACE/Praeger.

Linacre, J.M. 1989. Many-facet Rasch measurement. Chicago: MESA Press.

Linacre, J.M. 2002. Optimizing Rating Scale Category Effectiveness. Journal of Applied Measurement, 3, 85-106.

Linacre, J.M. 2009. FACETS (Computer program, version 3.66.1). Chicago: MESA Press.

Linacre, J.M., Wright, B.D. 2002. Construction of Measures from Many-Facet Data. Journal of Applied Measurement, 3, 484-509.

Lumley, T. 2005. Assessing second language writing: The rater's perspective. Frankfurt am Main: Peter Lang.

Lumley, T., McNamara, T.F. 1995. Rater characteristics and rater bias: Implications for training. Language Testing, 12(54), 54–71.

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.

MacMillan, P.D. 2000. Classical, Generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. Journal of Experimental Education, 68(2), 167–190.

McNamara, T.F. 1996. Measuring second language performance. London: Longman.

Messick, S. 1995. Standard of validity and the validity of standard in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5–8.

Myford, C.M., Wolfe, E.W. 2004. Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. In E. V. Smith y R.M. Smith (Eds.). Introduction to Rasch Measurement (pp. 460-515). Maple Grove, MN: JAM Press.

Park, T. 2004. An Investigation of an ESL Placement Test of Writing Using Many- facet Rasch Measurement, Papers in TESOL & Applied Linguistics, 4, 1-21.

Prieto, G. 2011. Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement. Psicothema, 23, 233-238. [Performance assessment using Many-Facet Rasch Measurement].

Razak, N., Khairani, A.Z., Thien, L.M. 2012. Examining Quality of Mathemtics Test Items Using Rasch Model: Preminarily Analysis. Procedia - Social and Behavioral Sciences, Volume 69, 24 December 2012, Pages 2205-2214. https://doi.org/10.1016/j.sbspro.2012.12.187

Tyndall, B., Kenyon, D. M. 1996. Validation of a new holistic rating scale using Rasch multi- faceted analysis. En A. Cumming y R. Berwick (Eds.), Validation in language testing (pp. 39-57). Clevedon: Multilingual Matters.

Wolfe, E.W. 2004. Identifying rater effects using latent trait models. Psychology Science, 46(1), 35–51.

Wright, B.D., Masters, G.N. 1982. Rating scale analysis. Chicago: MESA Press.

Wright, B.D., Linacre, J.M. 1994. Reasonable mean-square fit values. Rasch Measurement Transactions, 8, 370-371.

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