Polat, M. (2020). A rasch analysis of rater behaviour in speaking assessment. *International Online Journal of Education and Teaching (IOJET)*, *7*(3). 1126-1141.
https://iojet.org/index.php/IOJET/article/view/902

Received: 14.05.2020
Received in revised form: 11.06.2020
Accepted: 20.06.2020

A RASCH ANALYSIS OF RATER BEHAVIOUR IN SPEAKING ASSESSMENT

Research Article

Murat Polat

Anadolu University

mpolat@anadolu.edu.tr

Biodata(s): Murat Polat holds a Ph.D. at Osmangazi University, Educational Sciences, Research Methods and Statistics Program. Currently he is working as a language instructor at Anadolu University, Foreign Languages School. His research interests include language assessment, alternative testing and educational statistics.

Copyright by Informascope. Material published and so copyrighted may not be published elsewhere without the written permission of IOJET.

A RASCH ANALYSIS OF RATER BEHAVIOUR IN SPEAKING ASSESSMENT

Murat Polat

mpolat@anadolu.edu.tr

Abstract

The assessment of speaking skills in foreign language testing has always had some pros (testing learners' speaking skills doubles the validity of any language test) and cons (many test-relevant/irrelevant variables interfere) since it is a multi-dimensional process. In the meantime, exploring grader behaviours while scoring learners' speaking skills is necessary not only for inter/intra-rater reliability estimations but also for identifying the potential stringent and lenient graders in the rater-group to act accordingly to settle the best matches for graders when paired-rater-scorings or cross-marking-gradings are preferred for increasing the objectivity. In this exploratory study, which was implemented in 2019, 6 expert speaking graders scored 24 English language learners' speaking interviews from their video recordings including an individual and a pair discussion task for each student. A Rasch model in which MFRM (Many Faceted Rasch Measurement) was utilised to explore the scoring behaviours of the expert graders in terms of stringency and find out if their grading habits significantly affect language learners' overall speaking performances. The results of the present research showed that graders had significant score differences among each other and some of them scored too leniently or too stringently that might affect learners' speaking grades significantly.

Keywords: assessment, reliability, foreign language testing, rater bias, Rasch analysis

1. Introduction

In foreign language assessment, it is common to carry out performance assessment tasks through writing and speaking tests in which generally human-raters are involved in the procedure to mark the spoken or written responses that learners create in the target language so that we can increase the validity of testing. In the end of this process, the scores raters assign through those language tests are used not only to identify students' foreign language levels, but also to make educational inferences about the quality of the performance tests, flow of the language curriculum, sustainability of the language outcomes and finally the benefits of the language materials. Therefore, language test scores have a number of important predictive functions on which language program designers and school administrations base their substantive educational decisions. That is why these scores (particularly the subjective ones which are given through speaking and writing exams) have to be reliable and reflect the actual language performance of the testees.

However, the concept of language testing is a multi-dimensional task and draws on a wide and diverse set of personal, cognitive, and linguistic qualities which are inter-related and have unique functions in demonstrating a person's foreign language proficiency (Taylor & Wigglesworth, 2009). To illustrate, in a foreign language reading test, students may be asked to read an English text and answer some comprehension questions which were prepared to identify the main idea of the text or some other questions to be able to check if the student could make some inferences based on the information s/he gets from the reading text. The main objective is to measure the language proficiency of the learner, and the tool is the reading test here, but what if the student had an outstanding reading mastery in his/her native language or

vice versa, what if the learner already had many problems in reading even in his native tongue? Would they have an impact on this student's reading test score in a foreign language test? Which feature in such a test do you think would be more prominent: the test-taker's overall reading ability or his/her reading ability in the target language?

In another example of testing, students could be asked to demonstrate their foreign language skills through writing an opinion paragraph or an argumentative essay in which they were supposed to write about their favourite free-time activity with a number supporting reasons and examples. Normally, these essays are mostly scored by human-raters using either a holistic or an analytic rubric to set a standard within him/her scorings and with the other raters as well if multiple scoring sessions are held. In this case, a testee's receiving a high grade from this essay might depend not only on her/his foreign language mastery and the quality of writing, but also on the personal traits of one of the juries who grade the exam paper, such as the rater's habit of scale-shrinking (using a particular part of the scale and assigning similar scores regardless of the superiority of the written performance), lenient or harsh scoring (Fulcher, 2003). For example, think of a female rater who hates football and would never bother seeing or reading something about it or a male rater who loves watching football and reading comments about important matches They both grade the same student's essay on how amusing it is to watch a derby football match and share this joy with some friends. Do you think that these two graders would never be affected from their personal feelings while scoring this paper although they are both experienced, trained and were given the same grading criteria? Or what if one of them was a lenient, the other was a stringent rater? Would it somehow cause a scoring difference if they score this paper as a pair?

The third example could be about a speaking interview where two raters grade a paired speaking test in English. Let's assume that each grader was given a different task (welcoming the students and ensuing the identity check, employing the first or the second part of the test, delivering or presenting the exam questions etc.). In this case, the number of variables that could influence the students' scores might be more than the variables that might have an effect in a writing test. The tone and the body language of the graders, their way of asking the questions (stress, pauses, intonation etc.), students' familiarity, their reactions when the speaker does not understand a part or the whole question, how they listen to the students' responses (and the things they do meanwhile), their psychological moods at the time, expected test duration, the number of the testees (it is sometimes more advantageous to be in the first or in the last group of examinees in terms of receiving higher grades) may all have impacts on the overall speaking scores (Wang Haizhen, 2008). Even the difficulty of the exam questions might differ a lot, sometimes a jury's interview questions could be much more difficult than the other, and this can be quite occasional in exams where hundreds of students are interviewed at a time. Moreover, in paired grading sessions, where two different graders are assigned to grade students' speaking performances and compare their scores in the end of each grading session where rater negotiation is required, could it be possible that the experience, rank difference or the dominance of a rater would affect the other rater's scores when there are significant score discrepancies between raters?

All those examples might prove the fact that foreign language assessment (either in writing or speaking) have various dimensions (which are known as facets in testing terminology) which require further planning, research and extensive discussions to be able to design more bias-free and objective measurement tools and techniques. This study aims to explore rater behaviours in speaking assessment according to various independent variables; therefore, the findings of this study could be valuable for foreign language test designers since it aims to reflect the comparisons of rater judgements and score means according to various facets including the most debated topics such as students' language competence levels, students' individual

performance differences, grader differences and the differences which stem from the rubric's components. Moreover, the results of this research could also help raters re-check their individual scoring habits and rituals since those individual differences might affect not only a student's foreign language competency score but also his/her academic carrier in a bigger picture.

1.1. Theoretical background

1.1.1. Speaking Tests' Facets

In speaking interviews where the number of independent variables that might affect the flow of the exam, test-takers' moods and graders' judgements are so many, it is the duty of test designers to foresee such distractors and take necessary measures to minimise the test irrelevant factors (Brown, 1995). The first example in the introduction part, the one related to the reading comprehension test, illustrates a commonly encountered case in foreign language tests related to two relevant facets in measurement: testing the language skills and cognitive skills (directly or indirectly) at the same time. From one perspective, each student is a single facet, and each response of this student is an element of the item facet which is valuable and should not be considered separately from the student (Pollitt & Murray, 1996). Therefore, it could be impossible to distinguish cognitive skills of a learner from his/her foreign language skills while answering questions because each answer is unique and specific to its owner (Eckes, 2009). From another perspective, Wigglesworth (1993) reminded that the duty of test-designers is to focus on the target language outcomes and maximize the assessment of skills which are directly related to language production skills. Thus, the consideration of the language proficiency level of the students while preparing test items and designing the rubric according to such levels are critical in order to make fair judgements (Cohen, 1994). Considering this fact, the examinee and his/her language proficiency level constitute the first facet of the speaking tests.

Next, comes the role of the interviewer or the grader among the considerable facets of speaking tests. The second example in the introduction part, the one related to two different raters (the female rater who hates and the male rater who loves football and had to score a student's essay about his love for watching football), illustrates a commonly encountered case: rater-factor in assessment (Engelhard, 2002; McNamara, 2000). Research on rater judgements in foreign language assessment has revealed the fact that subjectivity of a rater is sometimes inevitable and judgements of even the most experienced raters may even vary significantly from the others' (Shohamy, 1983; Fulcher, 2003). Lumley (2002) stated that even the interpretation and application of a standardized scoring rubric may vary significantly and might cause significant scoring differences. The use of adjectives in descriptors may cause interpretation differences; for example, "extensive use of vocabulary items" or "a good control of cohesive devices", in this case, it might not be clear to identify "*extensive*" or come to an agreement on "*good control*". Another study carried out by Lumley and McNamara (1995) revealed that even extensive rater-training may not help to eliminate the grader differences in writing and speaking assessment in foreign language testing since raters reflect their prior-experience, prejudices, habits and beliefs to their judgements. What is more, in most of the cases where raters use their own judgements rather than using the grading rubric's descriptors, they are not aware of doing this and not accepting the fact that they score subjectively regardless of what he/she is supposed to do in the exam (Milanovic et al., 1996; Orr, 2002). That is why the rater behaviour in using the scoring rubric and its components is another facet in speaking tests.

Another issue which is worth mentioning is the complexity of a face-to-face interview, which is hard to handle both for the graders and the students for various reasons under normal circumstances. As illustrated in the third example in the introduction part, there are various factors that must be controlled simultaneously. These variables such as the tone and the body language of the graders, their way of asking the questions (stress, pauses, intonation etc.), test-takers' familiarity, grader reactions when the testee does not understand question, graders' way of listening the students' responses (and the things they do meanwhile), their psychological moods at the time which might have invisible and easily unrecognisable links in the background (Lane & Sone, 2006). These variables including students, test-tasks, graders, scoring rubric, and the atmosphere all together have effects on determining the speaking scores of the language learners' (Bachman, 2004). As all these cases illustrate, when subjective measurement is involved in foreign language assessment, it is inevitable that some human features will come out either willingly or unwillingly (Linacre, 2002; Wolfe & Dobria, 2008), and the speaking exam raters' leniency and stringency in scoring is one of these features.

Wolfe (2004) claimed that in some cases raters are, just because of their characteristic features, too lenient or too harsh in grading. The former could be considered as a positive attitude by the test-takers; however, regarding the consequences of scoring exaggeratedly high, lenient scoring would even harm the validity of the testing process deeply. Also, Myford and Wolfe (2004) underlined the same fact and warned that leniency in grading could be a serious measurement error when students are ranked according to those grades. Congdon and McQueen (2000) asserted that "rater stringency or severity" is the likelihood of giving low marks by a rater whose expectations are above other raters and tend to assign lower grades, which has been a phenomenon in testing for many years. Some teachers are known to be stringent graders whereas other raters are very comfortable in assigning high grades which might turn the assessment process into heads or tails, and if the rater is lenient you hit the jackpot, you lose if it is a stringent grader regardless of the quality of your performance (Lunz et al., 1990). In other words, the testee fails or passes no matter if he/she deserves or not). Al in all, defining the severe and lenient graders in the rater-pool is essentially important in quality and reliable scoring where cross marking assessment practices are held. Finally, the raters' leniency or stringency while grading students' spoken performances constitute the third facet in speaking assessment.

1.1.2. Many Faceted Measurement in the Assessment of Speaking

It is possible to categorize the theories used to analyse test results under two specific categories as Classical Test Theory (CTT) and Latent Trait Models (LTM). Within LTM, which was developed as an alternative to CTT, there are two separate models called as "Item Response Theory" and "Rasch model". Linacre (1993) developed the Many-Faceted Rasch Measurement Model by adding the scorer's stringency/leniency facet to the model developed by Rasch (1980) (Talent Levels of Individuals - Difficulty Levels of Questions). Many-faceted Rasch estimation alludes to the utilization of a class of estimation models that target giving a detailed investigation of different factors conceivably affecting the language test or its evaluation results (Kubinger, 2009). What is more, Di Nisio (2010) stated that MFRM has a number of superior properties compared to conventional measuring methods. To exemplify, the Rasch model uses measurement values of individuals (free from measurement errors) instead of raw scores.

In this model, each grader is assumed as a distinct facet which allows the researcher to explore probable scoring variety by investigating interactions of other graders with the other facets in the same grading process. To illustrate, Schaefer (2008) noted that the Rasch Model could be used to facilitate a considerable degree of rater objectivity in speaking or writing assessments by investigating the level of rater-bias. In addition, it was stated that Many Faceted Rasch Measurement (MFRM) could function as a powerful and substantial analysis in speaking tests since it can be helpful in detecting the measurement errors or sources of variance on students' test scores besides other variables such as item difficulty or test-taker's actual performance (Engelhard, 1992). In speaking assessment settings where graders use analytic rubrics, Weigle (1998) proposed the use of Rasch analysis to investigate rater bias since this model analyses rater behaviours and pinpoints the cases when significantly severe or lenient scores are assigned.

While grading students' written or oral performances, McMillan (2000) recommended the application of correct Rasch Models to investigate the rater impact on scores to sustain concurrent validity of the language test. Therefore, detecting which graders score more leniently or harshly than the others in the rater-pool MFRM is an effective method. Likewise, Park (2004) and Di Nisio (2010) suggested the use of MFRM to analyse the scoring differences of the raters according to different components of the analytical rubrics since the Many-Faceted Rasch model also determines and establishes the rules of a linear connection between each surface in a research (for example, in this research, students' language levels, the quality of speaking performances, the components of the criterion used to evaluate students' performance, and the scoring behaviours of the raters were investigated). In short, the Many-Faceted Rasch model standardizes the surfaces by combining the surfaces in a common plane to achieve an unbiased and effective measurement, and offers the ability to compare individuals' ability to perform the task, the difficulty/ease of questions, and scoring leniency or stringency of the raters at the same time (Hubbard et al.,2006).

In his research, McNamara (2000) found statistically significant variations in rater behaviours while grading test takers' language performances by MFRM including the score variance between what graders thought they were performing and what they actually performed. In another research related to language assessment, Koizumi et al. (2019) studied grader-behaviour impact on language learners' scores using the Rasch model which emphasizes inter-rater reliability in foreign language assessment including the effect of various components in the rating criteria. However, in Turkish context, MFRM has been used mostly in program evaluation studies rather than defining grader behaviours in performance assessment (Semerci, 2011; Uyanık et al., 2018). Thus, this study aims to analyse rater behaviours in language assessment according to a number of independent variables including rater differences, students' language level differences and scoring rubric's components. Thus, this study aimed to answer the research questions below to probe this underexplored but critical matter in Turkish context:

1. Do experienced graders' scores differ significantly although they use the same scoring rubric?
2. Do experienced graders' scores differ from the others' in terms of grader-leniency or stringency?
3. Do experienced graders' scores differ significantly according to different components of the scoring rubric?

2. Methodology

This exploratory study was carried out in 2019 in a foreign language preparatory school of a state university in Eskişehir. The official permission necessary for the study was taken from the language school's administration after reporting them the aim and the scope of the study.

2.1. Participants

Including the grader group and the testees, there were two separate groups of participants in this research. 6 English language instructors (4 female, 2 male graders) who were working in that language school voluntarily participated in the study. All the raters held MA degrees in ELT and had been grading students' oral performances for at least 10 years. As for the students, 24 students (15 female and 9 male students aged between 18-21) from 4 different language levels (from A level to D level, in this language school A level is considered the highest language level whereas D is the lowest, 6 students from each level) agreed to give permission for the use of their speaking exam video recordings to be used in this study. Speaking interviews are held in the school as paired interviews and each pair of students' oral performance is scored by two raters.

2.2. Instruments

There were two main instruments in this study; the videos of the speaking exam and the analytic rubric which was used to score students' interviews. All the video recordings belonged to the same proficiency exam's speaking interview section. The speaking exam in this school has two parts. In the first section, students are asked individual questions and in the second part, they are supposed to have discussions on specific topics with their exam partners. The scorings were done by an analytic scoring rubric which has five components (content, grammatical competence, lexical competence, fluency and interaction). The analytic criterion was developed by the school's testing office and all the raters were quite familiar with the analytic scale since in all speaking exams, the same scale is used in the school for language assessment. The rubric's components range from 0-4 (0 and 1 stand for the poor and weak performances, 2 for average, 3 stands for good and 4 stands for the excellence in the related criterion in a single component. The maximum score is 20 (5 components x 4 pts.= 20 pts.) in this speaking interview.

2.3. Data Collection & Analysis

As it was mentioned before, this study was carried out in 2019 and all the participants contributed to the study voluntarily after they were provided the necessary information about the aim the study. To be able to control the interviewer effect, 2 raters from the voluntary rater group made all the speaking interviews of the 24 students. As mentioned before, paired speaking interviews are carried out in the assessment and each pair of students' oral performance is scored by two those two raters. The raters were not the sample group's teachers at school and they did not know that the students were from 4 different language levels, on the contrary, they thought that they were all students from the same language level. The same set of interview questions and the same analytic rubric were used in the interviews. When the interviews finished and 12 videos were ready, they were copied by the researcher and presented to each grader in CDs. All the graders scored the speaking interviews individually and presented the score charts to the researcher for analysis. The data set collected from the participants were computed and analysed using the FACETS (Linacre, 2002) program, which is generally prescribed for MFRM analysis to distinguish parameter estimations, vital inspection for conjoint estimation, examination of infit and outfit levels to get fit estimations of the dispersion appropriately.

3. Results and Discussion

While defining rater behaviours, MFRM model is an effective approach to determine rater-scores that are not in the normal distribution. Linacre and Wright (2002) suggest MFRM to be used for performance assessment cases in which a number of dependent or independent variables could be observed in the final grades such as language level difference, item difficulty, graders' scoring difference or score differences stem from rubric components which can ultimately cause serious measurement errors. The four faceted Rasch model presented by Rasch (1980) is,

"*log (Pnijlk /Pnijl(k-1) = Bn - Rj - Di – Tjk"*

In the formulae; Pnijlk stands for the possibility of item n scored as k by rater j; Pnijl(k – 1) is the possibility of item n scored as k – 1 by rater j; Bn stands for the speaking skill of the testee displayed through the interview; Rj is the possible stringency of the rater; Di is the level of item difficulty and Tjk is the difficulty of a single scoring rubric component comparative to other components. Although this equation is a simple linear addition model, logit scale is a negative value ranging from infinity to positive infinity. It is assumed that values in the logit scale are different for each surface.

Linacre and Wright (2002) recommended that the first step to utilise MFRM analysis is to scan the data set in terms of the quantity of the standardised values. It is recommended that less than 5% of z-scores (standardized values) in the data set (the score distribution according to 5 different components out of 4) should be equal to or more than 2, or less than 1% of the z-scores (standardized values) should be equal to or more than the critical level which is 3. It was found that out of 720 score entries (24x6x5=720) in this study 19 (2.63%) were equal or more than 2 and out of 720 score entries 5 (0.69%) were equal or more than 3 and those findings proved that the present data set was fit for the analysis.

In Figure 1, the variable map presents a general view of the analysis of the whole data gathered from students' speaking scores including the measurement scale (1st column), students' ranking (2nd column), rater severity (3rd column), proficiency levels of the students (4th column), difficulty levels of rubric components (5th column) and the score divisions according to the scale's scoring components (6th to 11 columns) respectively. All column names were identified from high-low, severe to lenient and hard to easy to make it easier to recognise the placement differences in the chart.

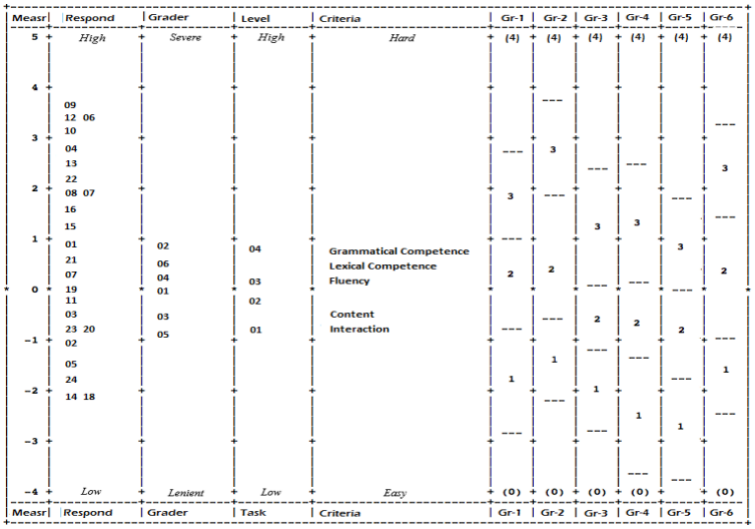

Figure 1: Vertical ruler map presenting the rank of students, graders, language levels and rubric components

Firstly, the student column (2nd column) in Figure 1 reveals the achievement rank of graders from the highest to the lowest score. The map shows that speaking score of student 9 (a participant from A level) was the highest and the first 4 ranks were all students form A level whereas the participants' coded as 14 and 18 (who were D level students) scores were the lowest ones in the score ranking. It should be reminded that the raters did not know the students' language levels and the findings presented in the 2nd column reveal that participants' achievement ranking is parallel with their placement levels at the language school. The third column shows the severity or leniency of the 6 graders via the scores they assigned in this research. Rater 2 (a female rater who had a 26-year-experience in language teaching and a similar expertise in language assessment) was found to be the most stringent rater whereas rater 5 (a male rater who had a 22-year-experience in language teaching and a similar expertise in language assessment) was observed to be the most lenient of the 6 graders. The results presented in the fourth column justified the findings in the second column. According to participants' language proficiency levels, Group 4, which stands for A level, had the highest scores whereas Group 1 (D level students) had the lowest scores from the rater group. As for the difficulty of the analytic rubric's components presented in the 5th column, grammatical and lexical competence components were found to be the most difficult components in the rubric whereas interaction and content components were the ones in which raters were more lenient in scoring. More detailed analysis of each facet will be presented in the following tables.

Table 1: *MFRM Measurement report according to language levels*

Obsv. Score	Obsv. Count	Obsv. Average	Fair Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Level
628	180	3.5	3.44	0.16	.07	1.1	1	1.1	1	4 (A)
574	180	3.2	3.17	0.06	.07	1.1	0	1.2	0	3 (B)
510	180	2.8	2.85	-0.08	.08	1.0	0	0.9	0	2 (C)
466	180	2.6	2.50	-0.14	.08	0.9	-1	0.8	-1	1 (D)
544.5	180.0	3.0	2.92	.02	.07	1.2	0.1	1.0	-0.0	Count:4
27.0	0.0	0.3	0.26	.12	.00	0.1	1.1	0.1	1.0	S.D.

RMSE (Model): .11 Adj S.D. .43 Separation Index: 1.92 Reliability: .81
Fixed (all same) chi-square: 18.2 df: 3 significance (p)= .00
Random (normal) chi-square: 3.1 df: 2 significance (p)= .39

The results which were presented in Table 1 reveal the total scores students from each language proficiency level received from graders (6 graders scored 6 students out of 5 different components in the rubric). The reliability coefficient in the Rasch analysis is 0.81. This result shows the reliability of the language level discrimination which could be identified as a reliable result since the reliability levels between 0.70 and 0.85 were classified as reliable values according to Eckes (2009). Similar to the reliability estimates for KR 20-21 or Cronbach Alpha tests, the computed reliability degree in Rasch analysis is the same with those analyses. Reliability level is a statistical value between 0-1 and it can be concluded that the higher the reliability level the better the analysis is. Considering the separation index 1.92 and the reliability coefficient 0.81 the null hypothesis about the score difference among students' language levels was rejected (χ_2 = 18.2, df = 3, p = 0.00). Thus, it means that there was a significant score difference between the language levels of the students ($p<0.05$).

MFRM has another superiority when compared with conventional methods in which it presents compared infit and outfit levels of the facets involved in the analysis. In the computation of these levels first, the chi-square test answers the question "Is there a statistical difference between the leniency or stringency of the raters?". Next, if the p value found for the test is less than 0.05, it can be interpreted that there is a significant difference between the

leniency or stringency of the raters while evaluating the students' performances. Thus, if there is a statistical difference between the raters, the raters who scored leniently or harshly can be determined from the detailed analysis (logit scale). The internal fit index (infit) is calculated according to the square statistics of the weighted standard residues, and the external fit (outfit) index is calculated according to the square statistics of the non-weighted standard residues (Engelhard, 1992). Since these fit indices show the difference between the expected value and the observed value with the minimum error, it is recommended to use the infit and outfit indices (Engelhard, 1992). According to Wright and Linacre (1994), the critical values for infit and outfit indices are between 0.6 - 1.4. Consequently, these indices are found according to the score differences estimated by the model and the scoring results of the raters. Thus, it can be concluded that infit and outfit statistics in none of the language level groups (Infit= 0.8-1.1, Outfit = 0.8-1.1) were below or above the critical limits, which means that there was no significant difference between the estimated scores and the students' assigned scores.

Table 2: *Speaking scores' measurement report (MFRM)*

Obsv. Score	Obsv. Count	Obsv. Average	Fair Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Student
117	30	3.9	3.75	0.25	.16	1.2	1	1.1	1	9
114	30	3.8	3.73	0.14	.16	1.3	1	1.2	1	12
114	30	3.8	3.70	0.12	.15	1.1	0	1.1	0	6
113	30	3.8	3.59	0.04	.15	0.9	1	0.8	-1	10
111	30	3.7	3.46	-0.03	.15	1.1	-1	1.2	0	4
108	30	3.6	3.42	-0.07	.15	1.0	-1	1.0	1	13
105	30	3.5	3.39	-0.10	.15	1.1	0	1.0	1	22
105	30	3.5	3.38	-0.12	.15	1.2	1	1.1	0	8
104	30	3.5	3.27	-0.19	.14	1.1	0	1.0	0	17
102	30	3.4	3.25	-0.22	.14	1.0	-1	1.1	1	16
99	30	3.3	3.18	-0.29	.14	1.1	-1	1.2	1	15
96	30	3.2	3.07	-0.36	.14	1.1	0	1.1	0	1
93	30	3.1	2.92	-0.39	.14	1.0	1	1.1	1	21
90	30	3.0	2.89	-0.41	.14	0.9	0	0.8	0	7
88	30	2.9	2.78	-0.49	.13	0.9	1	0.8	0	19
84	30	2.8	2.70	-0.54	.13	0.9	-1	0.8	-1	11
81	30	2.7	2.63	-0.58	.13	1.1	-2	0.9	0	3
79	30	2.6	2.48	-0.61	.13	1.0	1	1.1	1	23
75	30	2.5	2.34	-0.63	.13	1.1	1	1.0	1	20
72	30	2.4	2.21	-0.73	.13	0.8	-1	0.7	-1	2
69	30	2.3	2.15	-0.85	.12	1.1	-2	1.3	1	5
66	30	2.2	2.09	-0.89	.12	1.0	1	1.1	0	24
63	30	2.1	2.01	-0.92	.12	1.1	1	1.0	1	14
62	30	2.1	1.98	-1.11	.12	1.1	0	1.1	0	18
92.1	30.0	3.3	3.28	-0.48	.14	1.2	0.1	1.0	-0.0	Count:24
8.6	0.0	0.3	0.26	.38	.01	0.2	1.1	0.2	1.0	S.D.
RMSE (Model) .13 Adj S.D. .42 Separation Index: 2.81 Reliability .83										
Fixed (all same) chi-square: 82.4 df: 23 significance (p)= .00										
Random (normal) chi-square: 11.2 df: 22 significance (p)= .37										

The results which were presented in Table 2 reveal the scores each student received from graders out of 5 different components (6 graders scored each student out of 5 different components in the rubric). The reliability coefficient in the Rasch analysis is 0.83. Considering the separation index 2.81 and the reliability coefficient 0.83 the null hypothesis about the score difference among students' speaking scores was rejected (χ^2 = 82.4, df = 23, p = 0.00). Consequently, there was a significant score difference between the speaking scores of 24

students ($p<0.05$). The first 4 students' (participant 9,12, 6 and 10) mean scores in each component of the rubric was 3.8 which is considerably higher than the last 4 students' (participant 5, 24, 14 and 18) mean scores (2.2) in each component. When the infit and outfit statistics are checked, it can be concluded that infit and outfit statistics of none of the students' scores (Infit= 0.8-1.3, Outfit = 0.7-1.2) were below or above the critical limits, which means that there was no significant difference between the students' estimated speaking scores and the assigned scores.

Table 3: *Rater measurement report (MFRM)*

Obsv. Score	Obsv. Count	Obsv. Average	Fair Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Rater
445	120	3.7	3.32	6.25	.26	1.8	5	1.9	5	5
393	120	3.3	3.20	0.14	.08	1.1	1	1.0	1	3
382	120	3.2	3.06	-0.04	.07	1.1	1	1.0	0	1
357	120	3.0	2.87	-0.21	.07	0.9	-1	1.0	-1	4
336	120	2.8	2.66	-0.38	.07	0.8	-1	0.8	-1	6
276	120	2.3	2.52	-5.59	.06	0.4	-4	0.3	-4	2
364.8	120.0	3.1	2.94	-0.27	.07	1.0	0.1	1.1	-0.1	Count:6
23.1	0.0	0.3	0.33	.21	.01	0.2	.01	0.3	1.2	S.D.
RMSE (Model) .11	Adj S.D. .33	Separation Index: 2.99				Reliability .89				
Fixed (all same) chi-square: 248.2		df:5	significance: .00							
Random (normal) chi-square: 16.1		df:4	significance: .28							

Table 3 reveals the scores each rater assigned to 24 students' speaking performances out of 5 different components (24 students were scored out of 5 different components in the rubric). The reliability coefficient in the Rasch analysis is 0.89. Considering the separation index 2.99 and the reliability coefficient 0.89, the null hypothesis about the score difference among 6 raters was rejected (χ_2 = 248.2, df = 5, p = 0.00). In other words, there was a significant score difference between the judgements of 6 raters ($p<0.05$). It should be underlined once more that all the raters contributed to the study were all trained in speaking assessment and all had at least ten years of grading experience. The analysis revealed that Rater 5 was the most lenient one who gave 3.71 points out of 4 on average to each of the components in the analytic rubric; whereas, Rater 2 was the most stringent one who gave 2.30 points out of 4 on average to the same speaking performances. The infit and outfit results revealed similar findings. When those values for the graders' performances were checked, it can be seen that all the raters except Rater 5 and Rater 2 are within the accepted limits and it could be assumed that the other 4 raters (Rater 1, 3, 4 and 6) can assign closer scores to students' performances to the expected score ranges; however, Grader 5's infit (1.8) and outfit (1.9) and Grader 2's infit (0.4) and outfit (0.3) values are out of the critical limits (0.6 /1.4), which were reported by Wright and Linacre (1994). These findings lead us to conclude that Rater 2 and Rater 5 have significantly different scoring behaviours as Rater 5 is too lenient (3.71 score average out of 4) whereas Rater 2 is too harsh (2.30 score average out of 4) while scoring speaking interviews and these raters should not be paired with the other raters since their scoring difference threatens the reliability of the judgements of the overall rater-pool. A mean score difference of 1.41 out of 4 points between two raters is serious discrepancy despite the fact that both used the same rubric and both took the same trainings in norming sessions for scoring speaking interviews. Moreover, the leniency and stringency of these two graders change the whole score structure and this defect turns the entire score structure into an unreliable measurement in which there are too many outliers which are well above or below the acceptable limits.

Table 4: *Scale components' measurement report (MFRM)*

Obsvd Score	Obsvd Count	Obsvd Average	Fair Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Component
535	144	3.7	3.31	0.91	.06	1.8	5	1.9	5	Interaction
474	144	3.4	3.20	0.43	.08	1.1	1	1.1	1	Content
447	144	3.1	2.93	-0.08	.10	0.8	1	0.9	1	Fluency
419	144	2.9	2.73	-.31	.11	1.1	0	1.1	0	Lexic. com.
401	144	2.8	2.62	-.96	.12	1.0	-1	0.9	-1	Gram. com.
454.6	144.0	3.2	2.96	.08	.09	1.1	0.1	1.0	-0.1	Count:5
49.7	0.0	0.2	0.40	.48	.02	0.3	2.3	0.3	2.4	S.D.

RMSE (Model) .12 Adj S.D. .62 Separation 4.18 Reliability .80
Fixed (all same) chi-square: 276.4 df:4 significance: .00
Random (normal) chi-square: 24.2 df:3 significance: .29

The results which were presented in Table 4 reveal the scores assigned by 6 raters to students' speaking performances (6 graders scored 24 students). The reliability coefficient in the Rasch analysis is 0.80. Considering the separation index 4.18 and the reliability coefficient 0.80, the null hypothesis about the score difference among students' speaking scores was rejected (χ_2 = 276.4, df = 4, p = 0.00). Consequently, there was a significant score discrepancy between the judgements of 6 raters according to 5 separate components ($p<0.05$). It should be reminded that the analytic scale which was used in the study had 5 components (content, grammatical competence, lexical competence, fluency and interaction) and each was measured out of 4 points. The results also revealed that the mean scores assigned through "interaction" component were significantly higher than the other component scores which, at the same time, should be interpreted as these scores given in "interaction" component were not reliable. When the infit and outfit statistics are checked, it can be concluded that infit and outfit values of none of the components but "interaction" (Infit= 0.8-1.1, Outfit = 0.9-1.1) were below or above the critical limits. The "interaction" component's infit and outfit statistics (Infit= 1.8, Outfit = 1.9) reveal that raters were too lenient (3.70 score average out of 4) in scoring the oral interaction between students while speaking in the interview. The reason of this leniency might have stemmed from the lay out of the speaking exam. In one part of the exam, students have individual tasks in which the other student is not allowed to interfere even if he/she would like to speak. In the second part of the interview, a common task is given two both students and they are told to discuss the matter presented in the question which leads the students interact even if they don't want to, so measuring the quality of this interact in in the exam could be confusing and misunderstood by the raters. Therefore, the score means are exceptionally high in this descriptor.

Conversely, the mean scores of "grammatical competence" and "lexical competence" were significantly lower than the scores assigned for "content" and "fluency". This reveals another important concern in foreign language assessment in Turkey. The raters could be more critical while grading grammar and vocabulary since an important percentage of language teachers in Turkey still believe that just developing grammar and vocabulary knowledge and presenting this knowledge in oral or written performance is enough to succeed in language proficiency tests (Mirici, 2003). Namely, the average score difference between the two components "interaction" (3.70 score average out 4) and "grammatical competence" (2.79 score average out 4) is significant and is worth studying. Considering the fact that all the components in the rubric had an equal score distribution (4 points each), it might be interpreted that all those qualities were expected to be measured equally; however, the average scores assigned in those components revealed that grammar and vocabulary knowledge of learners were brought to the fore more, while the others were not taken into consideration sufficiently by the expert raters participated in this study.

4. Conclusion and Suggestions

This exploratory study, which was carried out in 2019 in a language school of a state university in Eskişehir, aimed to analyse rater behaviours in language assessment according to a number of independent variables including rater differences, students' language level differences and scoring rubric's components. For the purposes of the study, MFRM (Many Faceted Rasch Measurement) was used since each grader in this model is assumed as a distinct facet which allows the researcher to explore probable scoring variety by investigating interactions of other graders with the other facets in the same grading process. Finally, the results of the study revealed that there were significant score differences according to the language learners' proficiency levels and their individual performances. These findings were similar to the findings of McNamara and Ryan (2011) who emphasized the fact that even students from the same language proficiency levels might have been awarded very different scores since rater behaviours might change significantly while grading similar performances. Additionally, the results of the analyses revealed interesting findings in terms of leniency and stringency of the raters while grading the speaking interviews.

Rater measurement results revealed that there was a significant score difference between the judgements of 6 raters; even though all the raters contributed to the study were all trained in speaking assessment and all had at least ten years of language assessment experience. Ducasse and Brown (2009) mentioned about this finding in their study and concluded that regardless of their experience or expertise, raters prefer to make their own interpretations while scoring learners' oral or written performance because of many context-related reasons. The analyses also revealed that Rater 5 was the most lenient one who gave 3.71 points out of 4 on average to each of the components in the analytic rubric; whereas, Rater 2 was the most stringent one who gave 2.30 points out of 4 on average to the same speaking performances. Obviously, a 1.41-point-score difference out of 4 points in the mean scores between two experienced raters is a serious difference and such a problem should be meticulously examined and be explored why these two raters assign such different grades from the average distribution. Such a discrepancy among the raters in a rater group was also stated by Lumley (2002) and it was recommended that too lenient or too stringent graders should not be paired with the other raters who are scoring normally; instead, these raters should be invited to further training in language assessment and be retested before they core in the same rater-pool.

Finally, the analyses related to the rubric components revealed a significant score difference between the judgements of 6 raters according to 5 separate components including content, grammatical competence, lexical competence, fluency and interaction. The results revealed that the mean scores assigned through "interaction" component were significantly higher than the other component scores; whereas, the mean scores of "grammatical competence" and "lexical competence" were significantly lower than the scores assigned for "content" and "fluency". This finding reveals the fact that raters were more lenient while scoring the qualities such as interaction or the content of the oral production while they were more stringent while scoring grammar and vocabulary knowledge. This result leads us to see the fact that even the most experienced and trained raters overvalue the use of correct grammar and a variety of vocabulary items in assessing the speaking performances of language learners ignoring the importance of primary objective of speaking: communication. This finding was also reported by Shi (2001) and it was stated that non-native raters of spoken English care more on qualities like extensive vocabulary or complex grammatical structures and they score these qualities more stringently than do native raters of spoken English.

In conclusion, a number of suggestions could be made on the limitations of this study and related to the use of MFRM in analysing test scores. This study was carried out on voluntary

basis and 6 raters and 24 students participated in the study. Deeper and more reliable results can be taken from Rasch analysis models in which hundreds of students' performances are analysed by a bigger rater pool including more than 30 or 40 raters. Another suggestion can be made for testing units of language schools. The use of MFRM can be very helpful in defining stringent and lenient graders in the rater groups of language schools in the assessment of oral and written products and can serve well to re-train those raters, control and recheck their scoring behaviours and gain more reliable and fair exam results minimising the human effect to acceptable degrees. Eventually, in this study the researcher used some statistical methods and found a number of results using only quantitative data. A mixed method or a qualitative study could also be made on the leniency and stringency of raters and their reasons or feelings about (whether they know their lenient or stringent scoring and their reason of being so) this concern could be explored to have further information on the issue.

5. Conflict of Interest

The author declares that there is no conflict of interest.

6. Ethics Committee Approval

The author confirms that the study does not need ethics committee approval according to the research integrity rules in their country.

References

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Brown, A. (1995). The effect of rater variables in the development of an occupation- specific language performance test. *Language Testing, 12,* 1-15.

Cohen, A. D. (1994). *Assessing language ability in the classroom.* (2nd ed.) Boston, MA: Heinle & Heinle.

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

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

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26(3),* 423–443. https://doi.org/10.1177/0265532209104669

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), Tasks and criteria in performance assessment: *Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt, Germany: Lang.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Erlbaum.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Pearson Education Limited.

Hubbard, C., Gilbert, S., & Pidcock, J. (2006). Assessment processes in speaking tests: A pilot verbal protocol study. *Research Notes, 24,* 14–19.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model: Some critical suggestions on traditional approaches. *International Journal of Testing, 5,* 377–394.

Koizumi, R., Kaneko, E., Setoguchi, R., Innami, Y., & Naganuma, N. (2019). Examination of CEFR-J spoken interaction tasks using many-facet Rasch measurement and generalizability theory. *Language Testing and Assessment 8(2),* 1-33.

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. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3,* 85-106.

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

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to raters? *Language Testing 19/3*: 246-276.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12/1:* 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.

McMillan, 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. (2000). *Language testing*. Oxford, UK: Oxford University Press.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8(2),* 161-178.

Milanovic, M., Saville, N. & Shen, S. (1996). A study of the decision-making behavior of composition markers. In: Milanovic, M., Saville, N. (Eds.), *Studies in Language Testing 3: Performance Testing, Cognition and Assessment.* Cambridge University Press, Cambridge.

Mirici, I.H. (2003). The factors affecting the success in English proficiency exams and possible contributions of the internet. *Turkish Online Journal of Distance Education. 4(1)*: 1-8.

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.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System, 30/2*: 143-154.

Pollitt, A. & Murray, N.L. (1996). What raters really pay attention to? In: Milanovic, M., Saville, N. (Eds.), Studies in Language Testing 3: *Performance Testing, Cognition and Assessment.* Cambridge University Press, Cambridge.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainments tests.* Chicago IL: Mesa Press.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*, 465-493.

Semerci, Ç. (2011). The evaluation of students on ideas about the department of computer education and instructional technology (CEIT) according to Rasch measurement model. *5th International Computer & Instructional Technologies Symposium Proceedings.*

Shi, I. (2001). Native and non-native speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*, 303-325.

Shohamy, E. (1983). "Interrater and intrarater reliability of the oral interview and concurrent validity with cloze procedure in Hebrew". In J.W.Oller (ed.). *Issues in Language Testing Research. Rowley*, MA: Newbury House.

Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing, 26*, 325–339.

Uyanık, G.K., Güler, N., Teker, G.T., & Demir, S. (2018). Fen bilimleri dersi etkinliklerinin çok düzeyli Rasch modeliyle analizi. *Kastamonu Eğitim Dergisi, 27 (1*): 139-150.

Wang Haizhen. (2008). A Study on Raters' Interpretation and Application of the Rating Criteria in TEM4-Oral. *Theory and Practice of Foreign Languages Teaching 2:*33-39.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263-287.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10(3),* 305–319. https://doi.org/10.1177/026553229301000306

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

Wolfe, E. W., & Dobria, L. (2008). Applications of the multifaceted Rasch model. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 71–85). Los Angeles: Sage.