Pegem Journal of Education and Instruction, Vol. 12, No. 2, 2022 (pp. 213-225)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Comparison of Classical Test Theory vs. Multi-Facet Rasch Theory

Murat Polat[1*], Nihan S. Turhan[2], Çetin Toraman[3]

[1]Anadolu University, Eskişehir, Turkey

[2]Fatih Sultan Mehmet University, Istanbul, Turkey

[3]Çanakkale 18 Mart University, Turkey

Abstract

Testing English writing skills could be multi-dimensional; thus, the study aimed to compare students' writing scores calculated according to Classical Test Theory (CTT) and Multi-Facet Rasch Model (MFRM). The research was carried out in 2019 with 100 university students studying at a foreign language preparatory class and four experienced instructors who participated in the study as raters. Data of the study were collected by using a writing rubric consisting of four components (content, organization, grammar and vocabulary). Participants' writing scores were analysed thoroughly both by CTT and MFRM. At the first step, the participants' writing scores were calculated by taking the means of the writing points given by the graders in the CTT model. Then, the MFRM was applied to the data through a three-facet design considering the rater, student and rubric components as MFRM facets respectively. Finally, ability estimates obtained and reported in the logit scale via Rasch Analysis were converted into the analytic rubric's component scores used throughout the scoring procedure. Finally, two sets of writing scores were calculated and compared according to both measurement models. Considering the findings, it was summoned that there was a positive and high correlation between the ability estimates found according to the CTT and the MFRM. However, the mean score difference calculated according to both theories was still significant. Moreover, the analyses showed that criterion validity of the writing scores obtained via the MFRM was higher than the scores obtained via the CTT.

Keywords: CTT, Criterion validity, MFRM, IRT, Writing assessment.

Introduction

A remarkable number of important studies have been undertaken in order to attain the most valid and reliable measurement scores, identify the closest score to the real language performance (true score), discover the ways to have less workload and use less amount of resources (time, money and stationary) during the exams and the fastest and the most effective assessment models to implement production based, open-ended English writing tests more reliably and effectively. To start with, scoring writing papers have many challenges since the use of a standardized writing rubric, the presence of expert graders, well-planned writing tasks or ideal test conditions could not guarantee reaching true ability estimates (writing scores) for the exam papers. Writing tests require the use of productive language skills and as Lumley (2002) reported, foreign language students find the opportunity to express themselves better and express their thoughts freely with less stress and hurry in writing tests. In this respect, writing exams are excellent ways to test meta-cognitive skills such as analysing a problematic issue, organization, generating new and original ideas, evaluation, information use in different situations, establishing cause-effect relationships, generalizing, generating hypotheses and drawing comparisons between alternatives (Hamp-Lyons, 1995). Moreover, using writing tests have a number of superior advantages over multiple choice tests in testing productive skills since multiple-choice test items, in the end, include the correct answer which may absolutely harm the validity of the test itself. Thus, the first advantage of writing tests is that they can effectively measure some high-level skills that cannot be measured via multiple-choice tests easily (McNamara, 2000). Secondly, writing tests minimize the measurement error by eliminating the "chance factor" (Shaw & Weir, 2007). They are also suitable for partial scoring (Turner & Upshur, 2002) and that they can be prepared and administered more easily compared to multiple-choice tests (Kahveci & Şentürk, 2021)). Finally, in multiple-choice items, students have the chance to find the correct answer by elimination method; however, this is not the case for writing exams (Weigle, 2002). Accordingly, for test designers, presenting an item in the form of multiple-choice questions may be unsatisfactory to reveal the level to which a learner

Corresponding Author e-mail: mpolat@anadolu.edu.tr

https://orcid.org/0000-0002-9279-7699

How to cite this article: Polat M, Turhan NS, Toraman Ç (2022). Comparison of Classical Test Theory vs. Multi-Facet Rasch Theory. Pegem Journal of Education and Instruction, Vol. 12, No. 2, 2022, 213-225

Source of support: Nil

Conflict of interest: None.

DOI: 10.47750/pegegog.12.02.21

Received: 09.11.2021

Accepted: 28.01.2021 **Publication:** 01.04.2022

has acquired the intended educational goal which is aimed to be measured via an item in the test (East & Young, 2007).

On the other hand, besides the advantages, English writing exams have some drawbacks which are worth discussing. Initially, their implementation and scoring are time consuming and costly (Crusan, 2010; Karataş & Okan, 2021). What is more, it is difficult to ensure content validity in writing tests. Due to time constraints, the number of questions that can be asked in writing exams consisting of open-ended items is less than the item number in multiple-choice tests. Therefore, being limited to a small number of questions makes it difficult to provide full content validity in tests consisting of open-ended items (Andrade,1997). The last but not the least, the disadvantage of writing exams in terms of measurement is that they cannot be scored as objectively as they could be compared to multiple-choice tests (McNamara,1996; Romagnano, 2001). The score that students get in these exams may differ according to the person or people who assign that score. Therefore, the rater group (no matter how experienced or how trained they are) who take part in writing exams is a potential source of variance that might cause error in test scores (Tuzcu-Eken, 2021). Thus, it is necessary to explore how rater-based factors affect the measurement results while scoring the writing papers, that is, to determine the rater reliability (LeBreton & Senter, 2008).

There are a number of methods that can be used to determine rater reliability. Considering the main testing approaches, these methods can be classified under three headings; the ones based on the CTT, methods based on generalizability theory (GT) and the ones based on the MFRM (MacMillan, 2000). The measurement approach based on generalizability theory was not considered for this research as the measurement of English writing skills depend mostly on human factor and measurement differences which this reality would reveal. That is why in this exploratory study, the assessment-evaluation models based on the CTT and MFRM were particularly emphasized.

Literature Review

Using the CTT in writing assessment

The Classical Test Theory (CTT) is generally used in the measurement of foreign language skills due to its ease of use, popularity and easy analysis of the results. The CTT is built on the assumption that the observed score for a measure consists of the real (true) score and the measurement-error components. This assumption is illustrated by a simple equation as $X=T+E$ (X stands for the observed score, T represents the true score and E for the level of measurement error) (Engelhard & Stone, 1998). T or the real score is the arithmetic mean of the points that the individual will receive from a test if that test is administered to an individual an infinite number of times and the effects of learning problems or other unprecedented factors do not interfere in the measurement (Brown & Bailey,1984). However, since it is not possible to test a student's language skills an infinite number of times, the real score is an imaginary concept that is impossible to realize (Kline, 2005). E in the equation represents the error component. Considering the scoring process; the negative factors related to the measurement tool such as having weak items, item insufficiency; errors that are involved due to individual factors such as fatigue, anxiety, distraction, or factors related to the application of the measurement tool, unclear instructions and insufficient time all together constitute error variance. One of the factors that cause this error variance in tests that cannot be scored objectively is the factors related to the raters. In CTT, while determining rater reliability, different techniques such as rater agreement, Kappa statistics, Pearson correlation coefficient and comparison of means can be used (Vaughan,1991).

The level of agreement is the percentage of agreement among the raters, and it is obtained by dividing the assigned scores with which the raters agree by the number of writing scores (Davidson & Lynch, 2002). A second method which is commonly used to determine rater reliability in CTT is using the Kappa statistics. Although it is similar to the correlation coefficients, it differs in terms of its prior inclusion of the fact that some of the agreement among the raters is due to random chance and applying a correction process to control this agreement does not change the measurement error (David, 2008). Another CTT-based technique, used to examine rater reliability is the analysis of Pearson Correlation Coefficient. This coefficient level is a mathematical value between -1 and 1 representing the consistency of scores of the two raters' individual performances in terms of the trait they are to measure (Jones & Inglis, 2015). The last but not the least, another technique frequently used to determine rater reliability in writing assessment via the CTT is the comparison of rater-score means. When comparing those means, (if there are two raters, paired sample t-test is used) if there are more than two raters, analysis of variance (ANOVA) is used for repeated measures (Brown, 2012).

Using the MFRM in writing assessment

The Multi-Facet Rasch Model (MFRM) is a modern approach that tries to mathematically reveal the relationship between individuals' unobservable skills in a particular field and their responses to test items aimed to measure their abilities in the related field (DeMars, 2010; Kane, 2013). According to the MFRM, while the related parameters of a question (item) can be obtained independently of the respondent group who answered it, ability levels of individuals can also be estimated independently of the item sample in the applied test (Eckes, 2011). The MFRM consists of three separate models which could be listed as follows: a parametric model involving the item difficulty parameter; two-parameter model including

item difficulty and discrimination parameters, and a three-parameter model including item difficulty, item discrimination and chance parameters (Engelhard, 1992). However, a parametric model, also called the Rasch model, is the most basic model of the MFRM (Myford & Wolfe, 2003). The Rasch model was first developed for two-dimensional (1-0) measurement tools that can be scored as true/false or else (Laming, 2004). Later, different extensions of the basic Rasch model have been proposed for educational assessment, such as the partial scoring model discovered by Andrich (1978). Next, Masters (1982) proposed and developed the ranking scale in Rasch models. One of the other extensions of the basic Rasch model is the Multi-Facet Rasch Model (MFRM) introduced by the famous American statistician Linacre (2017).

Furthermore, the MFRM provides data on raters, scoring criteria, rubric, etc., as well as individuals' ability levels and difficulty levels of test questions. The MFRM is defined as a model that allows other sources of variability, such as test scores, to be considered in the evaluation (Knoch & Chapelle, 2017). In the MFRM, each source of variability with the potential to affect test scores is called a facet (Sudweeks et al., 2005). To illustrate, to test their English proficiency, university students are given a writing exam including 2 writing questions and the testees' exam papers are scored independently by 2 different raters. In this model, students' skills, writing questions and rater behaviour are sources of variability that can shape the scores. Therefore, in this case, we can talk about a three-faceted model in the form of student facet, item facet and rater facet. In the MFRM, the evaluation of factors related to raters as a facet that can cause variability in an individual's test score makes this model a well-fitting option for subjectively scored writing tests (Kobrin et al., 2011).

Classical Test Theory vs. Multi-Facet Rasch Theory

There are noteworthy differences between the CTT and MFRM in calculating ability estimates as well as in determining rater reliability in the evaluation of English writing skills. Baker (2001) argues that a raw score of a test taker calculated via the CTT is equal to the sum of the scores s/he gets from each item in the test. However, if a multiple-choice test is used in the assessment process, this scoring can be done automatically by a computer or a single rater. On the other hand, in writing exams consisting of one or more open-ended questions, using more than one rater in the evaluation process and calculating the arithmetic average of the scores given by the raters are methods commonly applied in order to obtain reliable estimates about the ability levels of students (Steedle & Ferrara, 2016). Accordingly, while calculating the ability estimates for the CTT in measurements made through open-ended questions, the scores obtained by the student from each question of the test should be added, this process should be repeated for all raters, and the arithmetic means of the test-points assigned by different raters should be calculated. However, the CTT does not provide information about how reliably individuals are distinguished from each other as a result of this process. In other words, while the reliability of the measurement results collected via open-ended questions is mentioned in the CTT, the reliability coefficients calculated for the test and raters are reported. On the contrary, no explicit information is provided on the reliability of the estimates made about the ability levels of students. By the use of the MFRM, on the other hand, in addition to the reliability of the items used in the measurement process and the scoring made by the raters, the reliability of the estimates about the ability levels of each student can be analysed (Goffin & Olson, 2011).

In The MFRM, all facets included in the MFRM are placed on a wide metric called logit in order to obtain capability estimates. While the CTT is doing analysis directly on raw points, in the MFRM, the measurements of each facet are converted to an evenly spaced logit scale. Then, considering the measurements of the components of other surfaces, estimates about the students' ability levels are made (Linacre, 1989). For instance, before calculations regarding the estimates of students' writing skills are made, the differences between the stringency/leniency of the raters are tried to be controlled statistically and the ability levels of the students are calculated by considering the differences in the scoring stringency of the raters and the statistical procedures applied to correct these differences (Elbow, 2012). Similarly; when reporting estimates of ability levels, difficulty levels of different items, if any, are also considered (Linacre, 2017). Accordingly, in the MFRM where there are three sources of variance in the form of student, rater and item, it can be said that ability estimates are obtained with the help of a function defined between the three facets processed in the analysis. Thus, ability estimates of students are calculated on all the scores given by all raters to all items (Elder et al., 2007). Considering these analyses calculating ability estimates in measurements made via open-ended questions with the CTT and MFRM, it is a critical question for assessment and evaluation researchers that what kind of similarities or differences exist between the ability estimates obtained according to these two theories.

Significance of the study and research questions

This research differs from previous studies which encompassed comparisons between the CTT and MFRM in terms of its purpose and scope including writing skill predictions. Therefore, the present study's results can be valuable especially for language institutions. It is also observed that there are studies focusing on the comparison of the CTT and MFRM in educational assessment models. In the study of Tobaş (2020), rater leniency and stringency were tested both by the CTT and MFRM, and no significant rater scoring difference was observed. Moreover, in another study reported

by Güler and Gelbal (2010), related to the comparison of CTT and MFRM, item and rater reliability levels analysed according to CTT and MFRM in open-ended questions were compared in detail. However, the comparison of ability estimates calculated according to these theories was not included in the study. In a study by Haiyang (2010), the CTT and MFRM were compared using an English test with open-ended items. However, the comparisons made were limited to rater and item reliability, as in the study of Güler and Gelbal (2010). In the studies conducted by Nalbantoğlu (2017) and Sudveeks et al., (2005) the results obtained from the CTT and MFRM and generalizability theory were compared in determining the rater reliability. Moreover, the research conducted by Huang et al., (2014) was limited to comparing the item difficulties, item discrimination and reliability values calculated in the CTT and MFRM. Thus, there are studies in the literature comparing CTT and MFRM in terms of rater reliability, item reliability, item difficulties and item discrimination. However, in the literature, very little has been reported on foreign language writing tests to compare the differences related to the ability estimates analysed considering the CTT and MFRM theories. In this respect, the present study can contribute to the literature on multi-dimensional ability estimation in foreign language tests and the consideration of different facets in the measurement of writing skills.

Since the ultimate goal of this exploratory research is to compare two different measurement theories, it is predicted that the study will also have an important scientific function. Considering the main goals of science which are to compare various theories, determine the functioning and non-functioning ones and to study the superiorities and weaknesses of these methods, findings and related discussions of this study, thus; are significant. With this aim in mind, comparing different assessment theories proposed for educational purposes, determining their strong and weak points and choosing the better and more practical one is regarded as a necessity for the coming studies in educational assessment and evaluation (Polat, 2020).

The aim of this study is to compare the writing-ability estimates analysed according to the CTT and MFRM. In order to achieve this, answers to the following research problems were investigated respectively:

1. What is the relative score-agreement between the writing-ability estimates analysed according to the CTT and MFRM in the assessment of English essays?
2. What is the absolute score-agreement between the writing-ability estimates analysed according to the CTT and MFRM in the assessment of English essays?
3. What is the criterion validity of the ability estimates calculated according to the CTT and MFRM in the assessment of English essays?

Methodology

This exploratory research was undertaken with language learners from a state university in Turkey in 2019. This model is an empirical approach that explores research questions that have not been studied in depth before (e.g., measurement of language skills, verbal or written performance assessment). In this type of research, in which a big pile of quantitative data is generally used, the variables of the design and the similarities and differences between a new or re-tested method within a large sample are examined. Due to its flexible and open-ended nature, this model was preferred for the research as it fits the motto of "Don't give up trying to reach the truth, which is the very basis of the scientific approach." (Greenberg, 1992).

Participants

The research was carried out with the voluntary participation of 100 students who received undergraduate foreign language preparatory education at university and four experienced instructors who collaborated as graders. Of the students aged between 17 and 21, 59 were girls and 41 were boys. Information on demographic characteristics of the raters in the study, such as gender, age, length of service in the teaching profession, and education level, was presented in Table 1. Since all of the raters graduated from an ELT department, Table 1 does not include an explanation with the type of undergraduate education of the raters.

Instruments

In this study, the ability estimates analysed according to the CTT and MFRM were obtained based on the English writing performances of the students. Therefore, the overall data of the study involved the scores of an English writing exam consisting of an opinion essay and an analytic rubric, which had been designed by the testing team of the language school, were used to score the writing skills of the students tested via this exam.

Writing test

The main goal of the study was to measure English writing skills, and other skills (speaking, listening, reading, etc.) that the participant students acquired while learning a foreign language were not dealt with. Therefore, the achievement test was developed specifically for this aim. Next, considering

Table 1: Rater descriptive info

Rater	*Gender*	*Age*	*Experience*	*Degree*
1	Female	38	15	ELT BA
2	Male	41	17	ELT MA
3	Male	45	21	ELT MA
4	Female	39	16	ELT BA

the accessibility of the raters who would evaluate the student papers, it was decided that the number of raters should be 4. Later, a writing test was prepared by the researcher including the following writing task: "University students must work for at least 3 months in a part-time job where they can earn their own money before graduation." Since the developed test would not be used to determine the academic achievement of the students, it was not deemed necessary to have the test examined by the experts in terms of content validity. However, the construct validity of the test was examined under the title of uni-dimensionality, which is one of the assumptions of MFRM. The measurement reports were presented revealing the reliability of the measurements obtained by the test for each facet in the output results of the MFRM analyses.

Analytic rubric

An analytical rubric designed by the testing team of the prep-school was used to score the essays written by the students in the writing test. In the rubric presented in Table 2, there are descriptors that fit 5 different success levels in 4 different components.

Data collection

The data of this exploratory study were collected in the Spring Semester of the 2018-2019 academic year. Before the application, the students were given a small briefing about the purpose of the research. It was stated to the students that the collected data would only be used for the purpose of the research and would not be shared with any other person or institution. The participants were told that the results of the research would not be used for grading purposes. However, in order to obtain valid and reliable results, the importance of writing the essays with the sensitivity of being in a real exam was emphasized. After the explanations about the purpose of the research were made, the students were reminded that there was no obligation to participate in the study. In this way, it was ensured that the research group consisted of only volunteer students. Students were given 60 minutes to answer the question in the writing test.

After collecting data from the students, the test papers were numbered. By doing so, the researchers aimed to prevent the raters from being affected by variables such as the student's gender and name during the evaluation. Since each of the

Table 2: Writing rubric

Scoring Rubric for Writing Skills			
Content	Perfect	5 Pt.	Covers the topic with a wide range of details (with necessary explanations and/ or examples)
	Good	4 Pt.	Presents some qualities of 5 and some of 3
	Average	3 Pt.	Covers the topic with a moderate amount of details (with limited explanations and/ or examples)
	Needs imp.	2 Pt.	Presents some features of 3 and some of 1
	Inadequate	1 Pt.	Fails to cover the topic with necessary details (with few & repetitive explanations and/or examples)
Organisation	Perfect	5 Pt.	Ideas organised well with a range of cohesive devices
	Good	4 Pt.	Presents some qualities of 5 and some of 3
	Average	3 Pt.	Ideas organised moderately with some coh. devices
	Needs imp.	2 Pt.	Presents some features of 3 and some of 1
	Inadequate	1 Pt.	Fails to organise with necessary cohesive devices
Grammatical Competence	Perfect	5 Pt.	A variety of gram. forms used accurately/appropriately
	Good	4 Ps.	Presents some qualities of 5 and some of 3
	Average	3 Pt.	Moderate variety of gram. forms used accurately
	Needs imp.	2 Pt.	Presents some features of 3 and some of 1
	Inadequate	1 Pt.	Inaccurate and inappropriate use of gram. forms
Lexical Competence	Perfect	5 Pt.	A variety of vocabulary used accurately/appropriately
	Good	4 Pt.	Presents some qualities of 5 and some of 3
	Average	3 Pt.	Moderate variety of vocab. forms used accurately
	Needs imp.	2 Pt.	Presents some features of 3 and some of 1
	Inadequate	1 Pt.	Inaccurate and inappropriate use of lexical devices

exam papers will be evaluated by four different raters, four copies of the papers were created by photocopying. Thus, the test papers are ready for the scoring process. After the scoring, writing skill estimates for the CTT were analysed by taking the means of the scores given by the four raters. Then, the same data were analysed according to MFRM, and ability estimates calculated in Rasch analysis were obtained for students' writing performances. Afterwards, the ability estimates calculated according to the CTT and MFRM were compared.

Data Analysis

Study data were the essays written by intermediate-level language learners, and these essays were scored by four experienced raters. The data obtained were analysed according to both the CTT and MFRM. For the CTT, in determining rater reliability, correlation coefficient results between raters and analysis of variance for repeated measures were used. The mean scores given by the four raters were calculated in order to obtain the ability estimates of the students' writing performance.

After the CTT analyses were completed, the analyses for the MFRM were done. At this stage, three facets were determined as rater, student and (rubric) component. Before performing the analysis, the assumptions of MFRM's uni-dimensionality, local independence, and model-data fit were tested. For the uni-dimensionality assumption, the average of the scores given by the four raters was taken and Exploratory Factor Analysis (EFA) was implemented over the calculated averages. Before EFA was performed, it was investigated whether the data set was suitable for factor analysis. In this respect, KMO and Bartlett tests were made. According to Tabachnick and Fidell (2007), the KMO value must be higher than 0.60 for the data to be suitable for factor analysis. In this study, the KMO sample fit coefficient was 0.62 and the Bartlett test value was 60.92 ($p < .001$). Accordingly, it can be said that the data were suitable for factor analysis. After this determination, EFA was performed using principal components factor technique. As a result of the factor analysis, a one-dimensional structure was reached, which explained 32.92% of the total variance and whose factor loads vary between 0.53 and 0.65. Accordingly, it could be said that the first assumption of MFRM, which is the one-dimensionality condition, was met. Also, for the variance rate explained in the EFA, considering that values above 30% was considered sufficient (Tabachnick & Fidell, 2007) and the value of .33 was accepted as the lower limit for the factor loadings (Çokluk et al., 2012). Thus, the construct validity of the writing test scores were also found to be sufficient.

Findings

In this section, findings of the research were presented considering the order the research questions were listed. First, students' writing scores were calculated according to the CTT. With this purpose, the means of the scores given by the raters were calculated and findings were presented in Table 3.

Table 3: Writing scores calculated according to the CTT

Student	*Score*	*Student*	*Score*	*Student*	*Score*	*Student*	*Score*	*Student*	*Score*
1	2,85	21	1.69	41	1.88	61	0.83	81	1.79
2	4.25	22	2.13	42	2.25	62	2.03	82	2.52
3	1.93	23	1.01	43	0.52	63	2.16	83	1.65
4	4.24	24	3.17	44	4.23	64	1.75	84	1.66
5	0.55	25	0.97	45	1.26	65	2.42	85	2.51
6	2.92	26	0.41	46	4.03	66	4.22	86	2.50
7	1.84	27	3.67	47	2.59	67	1.35	87	1.51
8	2.83	28	2.90	48	2.16	68	2.08	88	1.52
9	3.15	29	2.61	49	1.61	69	2.09	89	2.25
10	2.59	30	2.76	50	1.25	70	2.58	90	2.01
11	2.95	31	2.43	51	4.70	71	0.70	91	0.75
12	2.38	32	1.75	52	1.58	72	3.51	92	4.21
13	4.78	33	1.52	53	1.83	73	2.42	93	2.85
14	2.08	34	1.53	54	2.16	74	2.09	94	2.65
15	0.61	35	2.35	55	2.09	75	1.92	95	4.11
16	0.69	36	2.18	56	2.42	76	2.35	96	4.10
17	1.43	37	2.85	57	2.43	77	1.54	97	1.64
18	1.67	38	2.62	58	1.93	78	1.55	98	0.59
19	1.78	39	2.28	59	0.91	79	2.09	99	1.16
20	2.94	40	1.75	60	2.07	80	1.87	100	1.99
	Mean= 3.17					SD = 0.82			

As can be seen in Table 3, the ability estimates calculated according to the CTT for students' writing skills vary between 0.41 and 4.78. The mean and standard deviation values of the students' grades in writing skills were found 3.17 and 0.82, respectively. After calculating the ability estimates, inter-rater reliability coefficients according to the CTT was examined. For this purpose, first of all, the correlation coefficients between the raters were examined. The calculated correlation coefficients, along with the descriptive statistics of each grader's score means were shown in Table 4.

The findings presented in Table 4 revealed that correlation coefficients between raters ranged from 0.68 to 0.82, and all correlation coefficients were statistically significant. The fact that the correlation coefficients have values above 0.70 or very close to 0.70 suggests that the reliability between the raters is high (Bayram, 2009). However, since the correlation coefficient is a degree that is calculated independently from the mean scores and does not consider the absolute agreement between those scores, it is necessary to compare the graders' mean scores before talking about rater reliability. In this respect, the average scores of the graders were compared by applying variance analysis for repeated measures, and the related findings were presented in Table 5.

Checking the findings in Table 5, it could be concluded that there was a significant score-difference among the four raters [$F(3,97) = 105.91$, $p < 0.01$]. Considering the raters' average scores, the 3rd rater was scoring more generously than the other raters; thus, it can be said that 3rd rater behaved more leniently than other raters. Although there were strong correlations between the raters; the statistical significance of the analysis of variance results for repeated measures reflects that the absolute agreement between the raters is lower than the relative agreement.

After examining the ability estimates of students' English writing performances and inter-rater reliability level according to the CTT, a MFRM analysis was applied to the scores obtained. In the Rasch Analysis, first, the variable map, measurement reports of students, component and rater facets, and category statistics related to the rubric were presented respectively. Figure 1 shows the variable map reported as a result of the multi-surface Rasch analysis. In the figure students' ability levels, difficulty levels of the rubric components and, finally, the measurement units related to the stringency/leniency of the raters were represented. As can be understood, all sources of variability included in the analysis in MFRM are placed on a common scale called logit. In the 2nd column of Figure 1, the students are ranked in terms of their writing performance. The students' performance increases as they progress from the bottom to the up in this column. Therefore, participant number 13 had the highest writing performance. It can be said that the participant numbered 26 was the student with the lowest writing performance.

Obtaining measurements along the negative and positive ends of the variable map for students' writing performances reflects the successful differentiation of different writing performances apart from each other. In the 3rd column of Figure 1, there are measurements of the rubric components. In this column, where the rubric components were listed in terms of difficulty levels, the scores assigned in each component increases as you go from bottom to the top. Accordingly, component number 4 was the most striking one. It can be interpreted that the component number 4 (Lexical competence) was the one in which all the raters scored the lowest. It can also be seen that the component number 2 (Organisation) was the one in which all the raters scored the highest. The rubric components do not cluster at a single point, but located at different points of the variable map, and this means that students' performances on different rubric components can be effectively distinguished from each other. The 5th column of Figure 1 contains the measurements of the

Table 4: Inter-rater score correlation coefficients

Rater	*1*	*2*	*3*	*4*	*Mean*	*SD*	*Skewness*	*Kurtosis*
1	1				2.52	0.62	-0.06	-0.34
2	0.82**	1			2.07	0.45	-0.04	-0.12
3	0.72**	0.77**	1		2.96	0.71	0.36	-0.22
4	0.80**	0.78**	0.68**	1	2.68	0.69	0.33	0.55

**$p<0.01$

Table 5. Analysis of variance results on repeated measures

Rater	*Mean*	*SD*	*Wilks' Lambda*	*F*	*df*	*Error SD*	*Eta Squared*	*Test*
1	2.52	0.62	0.25	105.91	3	97	0.99	Significant score differences among raters
2	2.07	0.45						
3	2.96	0.71						
4	2.68	0.69						

raters. These metrics indicated that raters at the positive end of the column with higher logit scores were more stringent in scoring. It is also interpreted that raters who were at the negative end of the column and had low logit scores were more generous in scoring. Thus, the most stringent scores were assigned by the 2nd rater. It can also be said that the most lenient rater was the 3rd.

Although the variable map in Figure 1 gives important clues about the writing achievement levels of the language learners, difficulty levels of the rubric-components, and the strictness/generosity of the raters. In order to obtain more detailed information about student, item and rater surfaces, the measurement reports of each facet should be examined. Accordingly, the measurement reports for each facet were shown in this part respectively. First of all, the measurement reports of the student facet were examined and the obtained results were shown in Table 6.

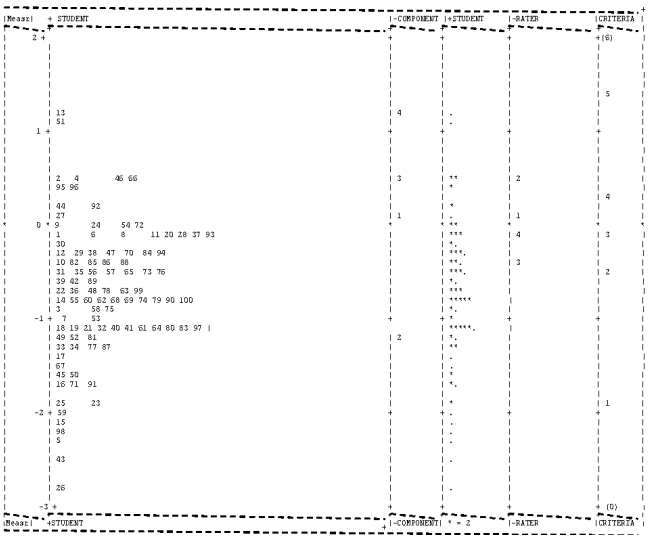

Fig. 1: Variable Map

Table 6: Writing scores obtained for the Student facet with MFRM

	Logit	*SD*	*Infit*	*Outfit*
Mean	-0.73	0.26	0.98	1.01
Std Dev.	0.75	0.05	0.47	0.64

RMSEA = 0.26 Adj S.D.= 0.70 Separation= 2.46 Reliability = 0.88

Fixed Chi-square = 586.8 df=99 p = 0.00
Random (normal) Chi-square = 82.4 df=98 p = 0.86

Table 7: Writing scores obtained for the Component facet with MFRM

Component	*Logit*	*SD*	*Infit*	*Outfit*
4 (lexical competence)	1.17	0.15	1.09	0.89
3 (gramm. competence)	0.81	0.13	1.16	1.07
1 (content)	0.59	0.12	1.01	0.98
2 (organisation)	0.21	0.11	0.99	1.02
Mean	0.00	0.07	1.07	0.99
SD	1.06	0.01	0.11	0.13

RMSEA = 0.06 Adj S.D.= 0.95 Separation = 13.00 Reliability = 0.98

Fixed chi-square = 1062.9 df=3 p = 0.00
Random (normal) chi-square = 4.0 df=2 p = 0.26

According to the findings given in Table 6, the average of the students' English writing performance is -0.73 and the standard deviation is 0.75 logit. Moreover, according to the results presented in Table 6, the averages of infit and outfit statistics were determined as 0.98 and 1.01, respectively. If the fit statistics are equal to 1, it is known that the fit between the model and the data set is acceptable (Tabachnick & Fidell, 2007). Therefore, the fitness statistics obtained for the individual surface reflect that the fit between the model and the data is acceptable. When the separation rate and reliability index values in Table 6 are examined, it is seen that the separation rate is 2.46 and the reliability index is 0.88. The high reliability index found in the analysis indicates that students with different writing performances can be successfully detected. In addition to the reliability index, the results of the Chi-square test show that students with different writing performances can be distinguished effectively. The results of the Chi-square test proven that there was a statistically significant score-difference among the students in terms of their writing performance [$\chi 2 = 568.8$, d f= 99, $p<0.01$].

After the measurements of the student facet, the measurements of the component facet were examined and the findings were presented in Table 7. The difficulty levels of the components vary between 1.17 logit and 0.21 logit, and a change of 0.96 logit was observed between the components in terms of difficulty levels. The mean for the difficulty levels of the components was 0.00 and the standard deviation was 1.06. In addition, the calculated infit statistics ranged between 0.99 and 1.16. It was determined that the outfit indices ranged between 0.89 and 1.07. The averages for infit and outfit indices were determined as 0.99 and 1.02, respectively. According to Linacre (1989), it is necessary to examine the fit statistics to decide whether there is an item in the analysis that adversely affects the fit of the model and the data. Güler

Table 8: Writing scores obtained for the Rater facet with MFRM

Rater	*Logit*	*SD*	*Infit*	*Outfit*
2	0.44	0.07	1.02	1.04
1	0.15	0.07	0.87	0.93
4	-0.16	0.06	1.05	1.03
3	-0.46	0.06	0.98	1.06
Mean	0.00	0.07	0.99	1.03
SD	0.36	0.00	0.07	0.08

RMSEA=0.06 Adj S.D. = 0.3 Separation = 6.04 Reliability = 0.96

Fixed chi-square = 152.5 df = 3 p = 0.00
Random (normal) chi-square = 3.1 df = 2 p = 0.25
Rater agreement = %51.42

and Gelbal (2010) stated that fit indices between 0.6 and 1.4 were acceptable. Considering this criterion, it can be said that there was no rubric component in the analysis that disrupts the compatibility of the model and the data. The fact that the averages of the fit statistics are very close to 1.00 means that the fit between the model and the data-set is acceptable.

The measurement reports of the rater facet were presented in Table 8. When Table 8 is examined, it is seen that the logit measures of the raters vary between 0.44 and -0.46. Thus, the logit range for raters' strictness and generosity is 0.9 logit. According to Table 8, the averages of infit and outfit indices correspond to 0.99 and 1.03 values. The fit indices are within the acceptable range of 0.6 to 1.4 for all raters. From this point of view, it can be said that there was no rater in the scoring that negatively affected the fit of the model and the data. Considering the separation rate and reliability index of the rater facet, the separation rate was 6.04 and the reliability index was 0.96. The separation level and the reliability index calculated for the rater facet showed a reliable difference between the raters, not a similarity in terms of scoring (Haiyang, 2010). Therefore, the calculated coefficients reflect that raters differ in their stringency and generosity in scoring. The significant Chi-Square values in Table 8 [$\chi 2 = 152.5$, df = 3, $p<0.01$] revealed that the difference observed between the raters was statistically significant. Finally, Table 8 shows that the absolute agreement between raters was 51.42%.

In the MFRM analysis, for each facet, outputs and category statistics are reported after the measurement reports. The category statistics obtained after scoring the essays in which students demonstrate their writing skills using an analytical rubric with a five-point rating were presented in Table 9.

To claim that the scoring rubric works effectively, there must be at least 10 observations in each sub-category of the scoring rubric (Lumley, 2002). The frequency values in Table 9 meet this requirement. Another indicator showing that the scoring scale works well is that as the categories of the scoring scale increase, the average measurements also increase (Linacre, 1989). The increase in the mean measurements in Table 9 in parallel with the rubric categories reveals that the scoring scale works effectively. The fact that the outfit statistics in Table 9 are very close to 1 is another indicator that reflects that the rubric works effectively.

Table 9: Category statistics of the rubric components according to the MFRM

Scores	*Frequency*	*%*	*Acc. %*	*Mean Measure*	*Expected Measure*	*Outfit Statistics*
1	254	16	16	-1.62	-1.64	1.1
2	437	27	43	-0.87	-0.82	0.89
3	540	34	77	0.08	0.05	1.08
4	298	19	96	0.83	0.83	0.93
5	71	4	100			

Table 10: Students' writing performance estimates calculated according to MFRM

Student	*Logit*	*Score*	*Student*	*Logit*	*Score*	*Student*	*Logit*	*Score*	*Student*	*Logit*	*Score*	*Student*	*Logit*	*Score*
1	-0.16	4.27	21	-1.15	2.99	41	-1.15	2.97	61	-1.15	2.97	81	-1.23	2.85
2	0.96	4.65	22	-0.68	3.57	42	-0.61	3.63	62	-0.82	3.38	82	-0.42	3.95
3	-0.90	3.25	23	-1.86	1.97	43	-2.67	0.62	63	-0.68	3.59	83	-1.13	2.97
4	0.94	4.64	24	0.12	4.41	44	0.92	4.62	64	-1.06	3.03	84	-0.28	4.07
5	-2.57	0.69	25	-1.33	1.33	45	-1.57	2.38	65	-0.48	3.83	85	-0.41	3.95
6	-0.08	4.35	26	-2.84	0.48	46	0.90	4.43	66	0.05	4.42	86	-0.41	3.97
7	-0.99	3.07	27	0.51	4.51	47	-0.34	4.01	67	-1.48	2.47	87	-1.31	2.73
8	-0.15	4.25	28	-0.08	4.33	48	-0.69	3.09	68	-0.76	3.45	88	-0.41	3.96
9	0.24	4.61	29	-0.34	4.01	49	1.08	4.38	69	-0.76	3.45	89	-0.62	3.83
10	-0.41	3.97	30	-0.21	4.19	50	-1.57	2.35	70	-0.34	3.99	90	-0.83	3.37
11	0.99	4.33	31	-0.48	3.83	51	1.15	4.80	71	-1.67	2.23	91	-1.21	1.53
12	1.16	4.43	32	-1.06	3.05	52	-1.22	2.86	72	0.38	4.59	92	0.51	4.54
13	1.17	4.89	33	-1.31	2.73	53	-0.99	3.05	73	-0.48	3.82	93	-0.15	4.17
14	-0.76	3.43	34	-1.31	2.75	54	-0.69	3.03	74	-0.76	3.05	94	-0.28	4.03
15	-2.35	0.71	35	-0.55	3.71	55	-0.76	3.43	75	-0.91	3.23	95	0.51	4.61
16	-1.67	2.25	36	-0.69	3.51	56	-0.48	3.82	76	-0.55	3.75	96	0.44	4.60
17	-1.39	2.61	37	-0.15	4.27	57	-0.48	3.83	77	-1.31	2.75	97	-1.14	2.97
18	-1.14	2.95	38	-0.34	4.01	58	-0.91	3.19	78	0.51	4.43	98	-2.47	0.70
19	-1.06	3.05	39	-0.62	3.69	59	-1.58	1.83	79	-0.76	3.47	99	-0.69	3.51
20	-0.15	4.15	40	-1.06	3.05	60	-0.76	3.45	80	-1.14	2.95	100	-0.83	3.33

On the other hand, in the MFRM outputs, measurements of all facets were reported in the logit table. This situation does not prevent the determination of the relative fit between the writing estimates obtained from the CTT and these estimates were all presented in the MFRM outputs. Similarly, the presentation of ability estimates in the logit scale in MFRM does not prevent the comparison of criterion validity of ability estimates calculated according to the two theories. Because while determining the relative fit and making comparisons in terms of criterion validity, the rankings made among the language learners in terms of their writing ability were taken as the basis and the unit in which the ability estimates were reported did not cause any difference on these rankings.

When Table 10 is examined, it is seen that the ability estimates of students' writing performances vary between 1.17 and -2.84 on the logit scale; in the scoring scale unit and it was seen that it extends between 0.48 and 4.89 out of a 5-point scale. After the ability estimates were converted from the logit scale to the units of the rubric, writing ability estimates calculated according to the CTT and MFRM were ready to be compared. Next, the correlation analysis was performed to determine the relative agreement between the ability estimates calculated according to the two theories and the dependent groups t-test was used to determine the absolute agreement presented in Table 11.

Considering the findings in Table 11, the relative agreement between the ability estimates calculated according to CTT and MFRM is extremely high [r = 0.97, p<0.01]. In other words, if a ranking was made among students in terms of their writing skills performance, the fact that the ability estimates were calculated according to the CTT or MFRM would not cause a difference in the actual ranking. On the other hand, the fact that the dependent group t-test results in Table 11 were significant [t (97), p<0.01] indicates that it is not possible to talk about an absolute agreement between the ability estimates calculated according to the two theories.

Table 11: Correlation coefficients & dependent-group t-test according to the CTT and MFRM

Model	*Mean*	*SD*	*N*	*r*	*t*
CTT	3.17	0.43	100	0.97**	54.83**
MFRM	3.29	0.51			

**$p<0.01$

While examining the criterion validity of the estimates of the writing skills obtained according to the CTT and MFRM, scores of the writing section of the students' end-course writing exams and proficiency exams in the same language school were taken as references. The calculated ability estimates and the correlation between these two variables were examined by correlation analysis, and the results were presented in Table 12.

The relationships between the writing skill estimates calculated according to both the CTT and MFRM, and the scores of the students' end-course exams with the writing section of the poficiency exams were statistically significant. Moreover, the correlation between the ability estimates obtained from the MFRM and the two variables taken as criteria (End-course & Proficiency writing exam scores) were higher. Accordingly, it can be said that the criterion validity of the writing skill estimates calculated in the MFRM was higher than the ability estimates obtained from the CTT.

Discussion & Conclusion

In this study; writing skills' estimates analysed along with the CTT and MFRM were compared to find out the relative-agreement, absolute-agreement and test-criterion validity. The findings from the research showed that there was a good relative fit between the ability estimates calculated according to the CTT and MFRM. This result, in case if a ranking is made among students in terms of ability estimates calculated on the CTT and MFRM, means that the rankings made according to the two theories could overlap with each other. In other words, if the measurement results driven from essay scores are to be subjected to a relative evaluation, the fact that the ability estimates calculated according to the CTT or MFRM may not affect the evaluation results. The same finding was underlined in a number of studies based on models in which the MFRM was used (Akın & Baştürk, 2012; Semerci et al., 2013; Yüzüak et al., 2015). This finding is significant since it has been revealed in many studies that one-dimensional measurement methods are not sufficient in the measurement of foreign language skills and those measurement methods that can measure different dimensions should now be used (Polat, 2020; Semerci et al., 2013; Zaman et al., 2008). For this reason, instead of the CTT,

Table 12: Criterion validity correlation coefficients

	1	*2*	*3*	*4*
1. CTT Performance measurement estimates	1			
2. MFRM Performance measurement estimates	0.97**	1		
3. End-course test result	0.66**	0.71**	1	
4. Proficiency test result	0.53**	0.59**	0.78**	1

**$p<0.01$

which has been used as the only measurement method in many language schools for a long time, it can be recommended to use IRT techniques, which allow doing more measurement by using less items.

Furthermore, the finding that ability measures calculated according to the CTT and MFRM rank individuals similarly is supported by studies showing that there is a high correlation between ability estimates obtained from different measurement theories. For example, in the study conducted by Zaman et al., (2008), it was found that there was a high correlation among the ability estimates obtained from the CTT and the two-parameter-IRT (Item Response Theory). Similarly, in the studies of Sims et al., (2020) and Eckes (2012), a high correlation was found between student achievement calculated according to the CTT and IRT. Again, in the study conducted by Schaefer (2008), high correlations were found between the ability estimates calculated according to the CTT and the ability estimates obtained according to the different models of the IRT. In the study conducted by Polat (2020), high correlations were found between CTT and success scores calculated according to single and multidimensional IRT. However, it should not be overlooked that these listed studies may indirectly support the research findings. Because while the listed studies were carried out on multiple choice tests, this study was carried out on open-ended items. Although both the one, two and three parameter models used in the mentioned studies and the MFRM used in this research are under the umbrella of the IRT, there are important differences for sure between these models.

Next, it was concluded that the averages of the ability estimates calculated based on the CTT and MFRM differ, and therefore, there is no absolute agreement between the ability estimates obtained according to these two theories. According to this result, it can be said that the evaluation results will differ in the case of an absolute evaluation based on the ability estimates obtained according to the CTT and MFRM. Considering that the average of the ability estimates reported in the MFRM is higher than the average of the ability estimates calculated in the CTT, a language student who is determined to fail according to the CTT can be found successful according to the MFRM results. Such a difference can lead to irreparable or very difficult consequences, especially for students whose scores are at the cut-off point (Brown, 2012). For example, when a language preparatory class student's score in the writing test is determined according to the CTT, it may be decided that the student has failed, and this decision may mean that the student might lose a semester and might, thus extending his/her faculty completion time. However, when the ability estimate according to MFRM is calculated for the same student, it can be concluded that the student got a score above the passing criterion and should pass the class. When the relative and absolute compatibility results between the ability estimates calculated according to the CTT and MFRM are considered together, it is inferred that in measurements made with open-ended questions, according to which theory the ability estimates are obtained will affect the absolute assessment results rather than the relative assessment. However, if there is a relative evaluation involving a threshold application, it should be noted that the evaluations made according to the ability estimates calculated in the CTT and MFRM may differ.

Another important finding driven from the research is related to the criterion validity of the scores analysed via the CTT and MFRM. In the criterion validity stage, the relationship between the ability estimates calculated according to the CTT and MFRM on the students' writing skills were examined. Considering the obtained results, it was concluded that the criterion validity of the writing scores analysed according to the MFRM were higher than the ability estimates obtained from the CTT. This result can be described by the fact that the MFRM is not based on the detection of rater differences, but to some extent controls the differences detected by the statistical corrections it applies (MacMillan, 2000). The findings of the criterion validity test proven that it would be more appropriate to calculate ability estimates according to the MFRM in measurements made for scoring writing skills.

As the last word, a number of limitations to the study and suggestions for further research could be communicated. First of all, it is a fact that the more realistic scores are obtained in studies where different measurement theories are compared, the more reliable and predictive results from the comparisons are. In this study, raters who grade students' essays may have acted more leniently in scoring because they knew that this was not a real exam, and this possibility may also have affected the decisions made. Next, "*the more the better*", the famous principle for the amount of data, applies to this study as well. Using more raters instead of 4, scoring much more student papers instead of 100 would undoubtedly be preferred in terms of obtaining more reliable and valid scores, but the number of participants was limited due to the voluntary design of this research. Considering those limitations, future researchers are recommended to apply similar patterns with different measurement techniques, especially in Turkish context and in different language schools, to model various other study-designs that would focus particularly on the measurement of productive skills (speaking & writing), and carry out studies with as many participants (both raters and students) as possible.

References

Akın, Ö. ve Baştürk, R. (2012). Keman eğitiminde temel becerilerin Rasch ölçme modeli ile değerlendirilmesi. *Pamukkale* Üniversitesi *Eğitim Fakültesi Dergisi, 31(*1), 175-187.

Andrade, H. (1997). Understanding Rubrics. *Educational Leadership, 54*(4).

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(9), 561-573. http://dx.doi.org/10.1007/BF02293814

Bayram, N. (2009). *Data analysis through SPSS in social sciences*. Ezgi Pub.

Brown, J. (2012). *Developing, using, and analysing rubrics in language assessment with case studies in Asian and Pacific languages*. Honolulu, HI: National Foreign Language Resource Centre.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*(4), 21–42.

Crusan, D. (2010). *Assessment in the second language writing classroom*. University of Michigan Press. https://doi.org/10.3998/mpub.770334

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için* çok *değişkenli istatistik SPSS ve LISREL uygulamaları*. Pegem Akademi.

David, A.B. (2008). Comparison of classification accuracy using Cohen's weighted kappa. *Expert Systems with Applications, 34*(2), 825-832. http://dx.doi.org/10.1016/j.eswa.2006.10.022

Davidson, F., & Lynch, B.K. (2002). *Testcraft: a teacher's guide to writing and using language test specifications*. Newhaven, CT: Yale University Press.

DeMars, C. (2010). *Item response theory*. Oxford, UK: Oxford University Press.

East, M., & Young, D. (2007). Scoring L2 writing samples: Exploring the relative effectiveness of two different diagnostic methods. *New Zealand Studies in Applied Linguistics, 13*(1), 1.

Eckes, T. (2011). *Introduction to Many-facet Rasch measurement: Analysing and evaluating rater-mediated assessments*. Frankfurt, Germany: Lang.

Eckes, T. (2012) Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behaviour, *Language Assessment Quarterly, 9*(3), 270-292, DOI: 10.1080/15434303.2011.649381

Elbow, P. (2012). Good enough evaluation: When is it feasible and when is evaluation not worth having. *Writing Assessment in the 21st century: Essays in honour of Edward M. White,* 303-325.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*, 37–64.

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171–191.

Engelhard, G. Jr., & Stone, G. E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement, 58*(2), 179–196.

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science, 6*(1), 48–60.

Greenberg, K. L. (1992). Validity and reliability issues in the direct assessment of writing. *Writing Program Administration, 16*(1–2), 7–22.

Güler, N., & Gelbal, S. (2010). Study based on classic test theory and many facet Rasch model. *Journal of Educational Research, 38 (1)*, 108-125. http://www.aniyayincilik.com.tr/main/pdfler/38/7_guler_nese.pdf

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analysing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics, 33*(2), 87-102. http://www.celea.org.cn/teic/90/10060807.pdf

Hamp-Lyons, L. (1995). Rating non-native writing: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

Huang, T., Guo, G., Loadman, W., & Low, F. (2014). Rating score data analysis by classical test theory and many-facet Rasch model. *Psychology Research, 4(*3), 222-231. http://www.davidpublishing.com/show.html?15856

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving. *Educational Studies in Mathematics, 89*(3), 337–355.

Kahveci, N., & Şentürk, B. (2021). A case study on the evaluation of writing skill in teaching Turkish as a foreign language. *International Journal of Education Technology and Science. 1*(4) (2021) 170–183.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Karataş, T. Ö., & Okan, Z. (2021). The powerful use of an English language teacher recruitment exam in the Turkish context: An interactive qualitative case study. *International Online Journal of Education and Teaching, 8*(3). 1649-1677.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation.* Thousand Oaks: Sage Publications.

Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing, Vol. 35*(4) 477–499 https://doi.org/10.1177/0265532217710049

Kobrin, J. L., Deng, H., & Shaw, E. J. (2011). The association between SAT prompt characteristics, response features, and essay scores. *Assessing Writing, 16*(3), 154–169. https://doi.org/10.1016/j.asw.2011.01.001.

Laming, D. (2004). *Human judgment: The eye of the beholder.* Thomson.

LeBreton & Senter, (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815-852. http://dx.doi.org/10.1177/1094428106296642

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

Linacre, J. M. (2017). *FACETS computer program for many-facet Rasch measurement*. Beaverton.

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

MacMillan, P.D. (2000). Classical, generalizability and multifaceted Rasch detection of interrater variability. *Journal of Experimenta Education, 68* (2),167-190. http://dx.doi.org/10.1080/00220970009598501

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. http://dx.doi.org/10.1007/BF02296272

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

McNamara, T. (2000). *Language testing.* Oxford: Oxford University Press.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Nalbantoğlu, Y., F. (2017). Analysis of the Rater effects on the scoring of diagnostic trees prepared by teacher candidates with the many-facet Rasch model. *Online Submission, 8*(18), 174-184.

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

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

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

Semerci, Ç., Semerci, N. & Duman, B. (2013). Yüksek lisans öğrencilerinin seminer sunu performanslarının çok-yüzeyli Rasch modeli ile analizi. *Sakarya* Üniversitesi *Eğitim Fakültesi Dergisi, 25*, 7-22.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing.* Cambridge University Press.

Sims, M., Cox, T., Eckstein, G., Hartshorn, J., Wilcox, M., & Hart, J. (2020). Rubric Rating with MFRM versus Randomly Distributed Comparative Judgment: A Comparison of Two Approaches to Second-Language Writing. *Assessment Educational Measurement: Issues and Practice, 39*(4),30–40.

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education, 29*(3),223.

Sudweeks, R.R., Reeve, S., & Bradshaw, W.S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261. http://dx.doi.org/10.1016/j.asw.2004.11.001

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics*. Boston, Pearson Education, Inc.

Tobaş, C. (2020). *Examination of the differential rater behaviours in performance evaluation with Many Facet Rasch Measurement. Gazi University Graduate School of Educational Sciences*. (Unpublished M.S. Master Thesis)

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49-70.

Tuzcu-Eken, D. (2021). Peer evaluation in writing: How to implement efficiently. *International Online Journal of Education and Teaching (IOJET), 8*(2). 708.

Vaughan, C. (1991). *What goes on in the raters' minds*? In L. Hamp-Lyons, (Ed.), Assessing second language writing in academic contexts (pp. 111–125). Norwood, NJ: Ablex.

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Yüzüak, A. V., Yüzüak, B. & Kaptan, F. (2015). Performans görevinin akran gruplar ve öğretmen yaklaşımları doğrultusunda ÇYRM ile analizi. *Eğitimde ve Psikolojide* Ölçme *ve Değerlendirme Dergisi, 6*(1), 1-11.

Zaman, A., Kashmiri, A., Mubarak, M., & Ali, A. (2008). Students ranking, based on their abilities on objective type test: Comparison of CTT and IRT. EDU-COM International Conference. http://ro.ecu.edu.au/ceducom/52/