Chabahar Maritime University
Iranian Journal of English for Academic Purposes ISSN: 2476-3187
IJEAP, 2021, 10(4), 86-103 (Previously Published under the Title: Maritime English Journal)

Examining Variability in EFL Writing Rating: The Role of Personality Types, Background, and Scoring Methods

[1]Abdolghafoor Bejarzehi
[2] Hossein Khodabakhshzadeh*
[3] Khalil Motallebzadeh

Research Paper IJEAP-2111-1801 DOR: 20.1001.1.24763187.2021.10.4.6.6
Received: 2021-10-16 Accepted: 2021-12-13 Published: 2021-12-15

Abstract

In today's world, effective written communication can be considered a necessary skill for English as a Foreign Language (EFL) learners. Furthermore, writing assessment has always been a challenging problem for teachers and raters. In the light of the foregoing, this research endeavors to examine the role of personality traits, rubrics, and scoring methods in the rating of Iranian EFL students' writings. To this end, data were collected from an initial sample consisting of 120 raters including both male and female IELTS instructors and EFL teachers teaching in English language institutes. Then 85 participants were selected based on the variability of their background to act as raters. A quasi-experimental and survey design was chosen to examine the variability and probable relations between and within the levels. This was done through a hierarchical model to investigate the relations beyond one level, thus including both a holistic rating scale and an analytic rating scale. The results revealed that multi-regression modelling was found the best model to overcome the deficiencies of the previous approaches. Besides, the findings showed that the multiple-trait scoring method specified more finely among the learners' writing capabilities, though both of these methods measured the same ability.

Keywords: Personality Type, Scoring Methods, EFL Writing, Variability, Writing Rating

1. Introduction

The assessment of EFL writing is such a demanding and time-consuming task that so many teachers might be tempted to neglect it in their instructional practice (Glušac & Milić, 2021; Peker & Torlak, 2020). Assessment is always laden with value, and it is inseparable from the identity of the authors and the unavoidable impacts of washback (Ahmed, 2018). Evaluating writing samples is a typical kind of language performance assessment (Huang & Han, 2013). Despite the ease of multiple-choice evaluation, the direct assessment of ESL or EFL learners' writings is both a complex and a demanding endeavor (Huang, 2010). Not merely do such factors as age, native language, culture, language proficiency, and task nature (Han, 2013; Han & Ege, 2013; Kormos, 2011) bring about variability in ESL/EFL learners' writing performances, but also there exist other sources of variation which affect the students' scores, i.e. essay characteristics, rating approaches, examiners' mother tongue, background, gender, and experience (Huang, 2012; Lim, 2011).

Although EFL/ESL learners might display different writing performances by nature, that part of variability in performance and scores which is created by raters and task types are not suitable since they cause errors in measurement and lead to unreliability in the writing scores (Huang, 2012). Examiners are an integral part of the writing performance evaluation; thus raters need to be trained and gain enough experience and expertise in order to improve their judgements (Lim, 2011). In a

[1] PhD Candidate in TEFL, abejarzehi@gmail.com, Department of English, Torbat-e Heydarieh Branch, Islamic Azad University, Torbat-e Heydarieh, Iran.

[2] Assistant Professor, kh.phdtbt2015@gmail.com, Department of English, Torbat-e Heydarieh Branch, Islamic Azad University, Torbat-e Heydarieh, Iran.

[3] Associate Professor, kmotallebz@tabaran.ac.ir, kmotallebz@gmail.com, Department of English, Torbat-e Heydarieh Branch, Islamic Azad University, Torbat-e Heydarieh, Iran. (Department of English, Tabaran Institute of Higher Education, Mashhad, Iran.)

nutshell, the said measures are the main sources of low reliability, lack of validity, and unfairness while measuring the ESL/EFL writing (Barkaoui, 2010a; Huang, 2012).

Previous investigations have mostly focused on factors relating to writing task types and raters' assessment behaviors with regard to the three main aspects of ESL/EFL writing assessments including reliability, validity, and fairness (Huang, 2012; Schoonen, 2005). Holistic and analytic evaluation approaches have been vastly practiced in writing assessment (Howard et al., 2020; Zhou et al., 2021). These two scoring approaches have both merits and demerits, depending on the task type and the consequences of the assessment practice.

Scoring and evaluation standards play an essential role in assessments which are mediated by raters. This is specifically more pronounced in cases such as multi-trait or analytic scoring approaches, where judgements are made in association with several measures developed to provide the core characteristics of the task under examination. Cognitively speaking, scoring measures channel the methods where examiners conceive and evaluate tangible samples of language performance and, finally, come to allocate scores to examinees (Lumley, 2005). Holistic scoring method enjoys the highest degree of construct validity; therefore, it is suggested as an instrument for awarding certificates, conducting placement, determining proficiency, and helping us in research examination (Weigle, 2002). Nonetheless, this method could pose threats to reliability because of its subjectivity owing to bias, tiredness, lack of internal consistency, students' background knowledge, and/or criteria changing from one essay to the next one.

On the other hand, the previous studies indicate that personality traits are of great importance in learning any language because the notion of language is naturally mixed with our feelings which will directly affect our personalities (Li et al., 2020). In addition, personality is defined as total-complex psychophysics of an individual which is influenced by environment, innate capacity, and several other intervening factors which are classified into two main types, i.e., extrovert and introvert (Yeung et al., 2012). Researchers have also asserted that there are different capabilities and understanding for students with these two types of personalities (Khodabandeh, 2021). As a result, this difference of personality helps learners represent different levels of achievement in their learning process, though their teacher is using the same teaching technique.

Reviewing literature, it can be concluded that there exists little research investigating the impacts of the scoring methods, personality traits, and rubric on the variability in Iranian EFL writing assessment context in one single study (Ahmed, 2020; Glušac & Milić, 2021; Khodabandeh, 2021). Hence, the focus in this research is on the role of raters' personality types, rubric, and scoring methods in a foreign language writing context. In fact, this study seeks to shed light on the scoring approaches and help English teachers working in similar situations with the decision-making processes that might let them enhance their writing assessment practice. Thus, this study aimed at investigating the following research questions:

Research Question One: Which background factors of the raters predict the score variability with regard to holistic and analytic methods?

Research Question Two: Which personality types predict the score variability related to holistic and analytic methods?

Research Question Three: How do the personality traits at the rater level affect the relationships between the scoring methods and the ratings at the rating level?

Research Question Four: How do the background factors at the rater level affect the relationships between the scoring methods and the ratings at the rating level?

2. Literature Review

This section covers the most relevant literature with regard to evaluating the EFL writing rating. For this purpose, the content is organized in two general sections, which includes the concepts and different trends regarding writing assessments and previous empirical studies.

2.1. Concepts and Different Trends Regarding Writing Assessments

As writing makes a bridge between thought and speech, it is seen as a discovery process that provides beneficial opportunities for ongoing learning. According to Vygotsky (1978), thought and speech are the human consciousness' essence, and writing is either speech in thought and image or, as Finnocchiaro put it, writing thinking (cited in Elahinia, 2004).

Writing is one of the most important and widely needed language skills for EFL students, allowing them to experience different actions (doing brainstorming, preparing drafts, revising the text, and editing it) where the writers are required to be actively revising their essays for further improvement (Elahinia, 2004). In other words, students are guided by the assessment feedback to analyze their works and make the required changes, taking necessary steps to express the meaning clearly. That is why assessment can result in language development for EFL writing students and hence assist their teachers (Lim, 2011).

For quite different goals, language teachers and raters perform language assessment widely around the world. Generally, language teachers make decisions on the basis of the garnered information using different assessment processes and instruments, and these decisions have undoubtedly certain consequences for main stakeholders, language institutes, institute managers, English teachers, and above all for students. According to Bachman et al. (2010), the primary usage of any language assessment is to gather information in order to make sound and fair decisions. These decisions have important effects and consequences in the educational and societal settings in which the assessment takes place, affecting various stakeholders, individuals and programs.

There are different perceptions of assessment among the experts in the field of language teaching. Pierce and O´Malley (1992), for example, emphasize the concept of authentic assessment as a classroom assessment's alternative method by pointing out that the alternative assessment includes any activity which helps us in understanding what a student knows or is able to do. This method is intended to show the students' improvement and guide their instruction; thus, it is a substitute to traditional testing policies, such as multiple-choice tests. Alternative assessment is by nature criterion-referenced and is typically authentic since it is similar to the activities that students usually do in their classroom and real-life situation.

The authors, holding this view, underline the necessity of establishing connections among classroom teaching, language syllabus, and assessment practices. Differentiating assessment from testing, Ahmed (2018) regards assessment to be a much broader process than testing. Assessment is by nature a never-ending process where the language instructor is continuously watching and making judgements on the performance of their students or the teaching process in the hope of comparing the previous performance with the present one, while testing involves a certain method to gauge a person's ability in a specified topic. In this regard, it is proposed that assessment and evaluation have similar meanings. These two concepts hold different conceptualizations for Bachman et al. (2010), though. They define assessment as collecting information about a specified topic resulting in a score or a verbal description, which is typically applied in evaluating the students (Bachman et al., 2010). Namely, when teachers make their most important decisions according to the assessments done in the classroom, the students are evaluated. Language teachers, while assessing, take an interest in judging the students´ or test takers´ level of proficiency, and this is done in order to reach certain assessment goals such as a) shortlisting and identifying, b) placement, c) putting students in new categories, d) checking learners' improvement, e) judging the efficiency of programs, and f) accounting for the established programs (Pierce & O´Malley, 1992). In this research, the meaning of assessment given by (Bachman et al., 2010) is adopted in this research.

There are five main ideas which should be taken into account in language assessment; a) practicality of the test, b) reliability of the instrument, c) its validity, d) authenticity level and e) washback effect. She adds that assessment should a) be clarified well for the students, b) follow certain established procedures, and c) aim to reach certain aims and objectives. Assessment's principles which are important in writing assessment are mainly the validity and reliability of the instrument, thus we need to take care of these issues (Peker & Torlak, 2020).

2.2. Previous Empirical Studies on Writing Rating Methods

Several empirical researches have explored how the rating methods yield the ESL/EFL writing ratings variability and reliability (Barkaoui, 2010a). Barkaoui (2010b), for instance, explored the impacts of holistic and analytic rating approaches on EFL learners' writing scores, rating procedure, and raters' notions. To reach the aims, four EFL writing instructors rated 32 compositions using both scoring scales. The data were collected through using think-aloud protocols while the raters evaluated two four-paper sets in a qualitative data analysis. However, *G*-theory approach was implemented in the quantitative data analysis stage. It was revealed that an exceptionally higher inter-rater reliability was found in holistic rating. Yet, in both rating procedures, the rating procedures were alike.

According to Weigle (2002), holistic scoring approach, on the whole, has fewer particular depicters than the analytic scoring approach. In the recent years, a scale with fewer depicters and a scale with more exact depicters for writing in an English-for-academic-purpose (EAP) situation were compared by Knoch et al. (2007); this was done to find out which scale could produce better results regarding validity and reliability. Implementing the two scales, ten experienced raters rated 100 papers. To compare raters' behavior, the quantitative data was analyzed using a multi-faceted Rasch measurement analysis, while the raters' perceptions of the efficiency of the two rating scales were determined through interviews and questionnaires. According to the findings, a significantly higher rater reliability index was observed in in the more-depicter scale than in the scale with fewer depicters.

Shabani and Panahi (2020) investigated the amount of agreement among the rubrics which are endorsed and used for evaluating the writing tasks in international tests of English. For this purpose, they randomly chose 200-essay samples written by IELTS candidates in an academic IELTS test. This sample came from a population of about 800 essays kept in an officially recognized IELTS center. The sample essays were written between 2015 and 2016. Their findings indicated that raters could greatly impact on the score of learners. Finally, the data analysis revealed a general agreement among the examiners and the scores given when correlation and factor analysis were used.

Zhang et al. (2015) made a comparison between two different scoring methods, the data coming from some three hundred essays assessed by fourteen examiners. The researchers analyzed the reliability indices of examiners for each method and investigated if the scores given by these raters were different. Findings indicated that the scores given were to some extent different when very many examiners were engaged in the process. Besides, it was reported that students possessing lower composition ability tended to receive higher scores in analytic scoring condition, while students who were more proficient in writing scored much higher when holistic scoring method was applied.

In the study conducted by Ghanbari and Barati (2020), the researchers conducted a study in order to develop and validate a new rating scale for scoring academic writing in the Iranian EFL context. They conducted their study in three different phases. In the first phase, authors conducted a semi-structured interview with several raters coming from different Iranian universities. In the second phase, they designed a questionnaire according to the points mentioned in the interviews as well as reading the relevant literature. In final Phase, the questionnaire piloted with 30 raters, who were experienced teachers teaching in the major state universities of Iran. Using FACETS to analyze the raters' performance. The researchers found that the new scale enjoyed high profile of reliability and validity.

Zhou et al. (2021), designed a method of pattern analysis which is appropriate for two-tier item pairs, this was done for Control-of-Variable (COV) of Lawson's classroom test of scientific reasoning (LCTSR). The needed data of the study were garnered from grade-4 students studying at USA and China colleges. Based on students' response patterns, the researchers came up with six performance levels, which could function as signs of COV reasoning improvement. They believed that their study provided fruitful information regarding the developmental levels of students' reasoning skills.

Ragupathi & Lee (2020) detailed the ways in which rubrics can be utilized as instructional as well as ranking instruments. They investigated the ways in which rubrics could assist teachers in giving useful feedback, hence upgrading the course design and clarifying the learning objectives.

Additionally, authors analyzed how evaluation can include qualitative and subjective judgments, thus suggesting that a good rubric can be a helpful tool in increasing fairness and consistency of ranking. Finally, implications were provided on how rubrics have the power to influence and improve our current teaching practices.

Conducting a study on the impacts of self-assessment and peer-assessment on improving the writing ability of Iranian EFL learners, Fathi and Khodabakhsh (2019) gathered data through administering two timed essay writing essays. These functioned as the pre-test and post-test. The analyses of the data revealed that both of the said assessment methods resulted in a significant improvement in the writing skills of the participants.

3. Methodology

This section is organized into main five parts, which discuss participants of the study, design of the study, instruments, the procedures, and data analysis.

3.1. Participants of the Study

The initial sample consisted of 120 raters including both male and female IELTS instructors and EFL teachers teaching in English institutes. After administering the background questionnaire, 85 participants were selected based on the variability of their background to act as raters in this study. All of them were Persian native speakers and their age range was 36-56 (Mean = 45.50). In this part, the purpose for choosing them from this sample of raters was to work with subjects that had a considerable amount of practical educational experience in a classroom setting. The final sample of the raters was selected from among the IELTS instructors from Sistan and Baluchistan, Khorasan, Tehran, and Fars provinces.

The number of IELTS candidates who were selected to write an argumentative essay was 32. They were selected from Foreign Language institutes in Chabahar, Tehran, and Mashhad. There were 17 men and 15 women, with an age range of 25 to 45 and similar first languages i.e., Persian and Baluchi, who had studied the writing task two in IELTS preparation courses and were ready to sit for the sample test according to their band scores on the writing task.

Stratified random sampling was utilized to select the participants both on the part of examinees and raters. As the sample size should follow the Multi-level Modeling (MLM) requirements in this regard, the number of the participants in Level-1 (IELTS Candidates) was 32 and the number of participants in Level-2 (Raters) was 80 selected from among 120.

3.2. Design of the Study

This study followed a quantitative approach through conducting survey to examine the variability and probable relations between and within the levels considered through a hierarchical model to investigate the relations beyond one level. Simply put, with regard to the data collection procedure, the present study employed both the quasi-experimental and survey designs.

3.3. Instruments

The following instruments were used to fulfill the requirements of the study.

3.3.1. Writing Samples

The writing samples were task 2 of IELTS test that is an essay of 250 words. The type of essay was argumentative one, so that all of the participants developed homogenized type of essay. The topic was also homogenized to avoid topic bias. They were asked to write at least 250 words. The topic (see appendix A) was given to the IELTS candidates to write their argumentative writing:

3.3.2 Analytic and Holistic Rating Scales

To obtain the research objectives and make fair judgements, assessments of the written compositions are better to be performed according to a standard rating scales and reliable assessment instruments (Weigle, 2002). To do so, the researchers made use of both holistic and analytic rating scales, the two most commonly and widely used rubrics (Rakedzon & Baram-Tsabari, 2017; Weigle, 2002). There were two reasons for including both of these scales. For one thing, it is quite usual for raters to use both of these rating scales in their evaluation of the writings, referring to the existing rubrics widely employed in the field.

While using the holistic rating scale, the examiner concentrates mostly on the work as a whole and pays attention to the main essential features of a piece of writing. Because of looking at the writing as a unit, the assessor makes a thorough judgement on the basis of several elements where loosely differentiated standards are put together as a whole and changed into a single quality. On the contrary, the analytic rating scale allows the examiners to evaluate the students' writing quality in accordance with some certain criteria, but this is done in a more rigorous, systematic, and differentiated manner. Thus, involving both of the existing rubrics improves the authenticity and practicality of the research. The second reason has to do with the research design, where this study examines the role of raters' personality types, rubric, and scoring methods in a foreign language writing context.

Based on a review of literature, the researchers developed some instruments for this study, consisting of both holistic and analytic scales (Weigle, 2002). Besides, Iranian EFL learners' essays, examiner and faculty input, and assessment objectives were taken into account for the development of the instrument. The holistic rubric was based on a ten-point scale addressing the following writing areas: a) grammatical accuracy, b) task content, c) organization and structure of the essay, d) writing style and quality of expression, and e) the mechanics of writing. It was also integrated with the descriptors in the analytic rubric. It is worth mentioning that the raters did not give equal weight to the specified categories, but the specified categories were given different weights. To be more exact, different point values were assigned for the five categories. In the following table, all the five weighted categories and the points belonging to each are shown.

Table 1: The Ten-point Scoring Criteria with the Assigned Weights (Han & Huang, 2017, p. 121)

Category	Weight Percentage
Grammar	30 %
Content	20 %
Organization	20 %
Style and quality of expression	15 %
Mechanics	15 %

3.3.3 Personality Trait Inventory

This inventory which is entitled Neuroticism-Extraversion-Openness Five-Factor Inventory (NEO-FFI) includes 60 statements and covers 5 main domains. This instrument is designed to be used in cases where time is an important factor, and the researchers have accessed enough general information about the personality type of the participants aging over seventeen years old (Costa & McCrae, 1992). Having made several necessary changes in the form and wording of the questionnaire items, Siyyari (2011) developed the Persian version of this questionnaire for the Iranian context. This newly localized version, used in the current research, has sixty items, and the respondents have to respond to the items based on a 5-point Likert scale. The questionnaire is reliable enough, with the reported reliability index of 0.80 (see Appendix B).

3.3.4. Background Questionnaire

In order to gather enough information about the background of the examiners, a researcher-made background questionnaire was administered to select the participants of the study from among the initial sample. It covered three main areas: personal background, professional training, and work experience (Appendix B).

3.4. Data Collection Procedure

Stratified random sampling was utilized to select the participants both on the part of examinees and raters. Some 32 IELTS candidates were selected to write an argumentative essay from Chabahar, Tehran, and Mashhad who had passed the writing task 2 training in IELTS preparation courses and were ready to sit for the sample test. The writing samples were task 2 of IELTS test that is writing an essay of no fewer than 250 words in 40 minutes. The type of essay was argumentative, so that all of the participants developed one type of essay. The topic was also the same for all the participants to avoid topic bias.

After selecting the IELTS candidates, the researcher arranged a session with the manager of Mahan institute, located in Tehran, to hold a mock exam of IELTS writing task 2 for the purpose of the study. A mock exam of IELTS writing task 2 was held and the researcher presented a topic (mentioned before), and the IELTS candidates were required to write at least 250 words in the allotted time. The reason was that this institute regularly holds mock IELTS tests, and there are many candidates registering there.

The initial sample of the raters consisted of 120 raters including both men and women IELTS instructors as well as EFL teachers teaching in English institutes. After administering the background questionnaire, only 80 participants were selected based on the variability of their background to act as raters in this study.

In order to increase the reliability of their scoring, all examiners in this research received a brief training session before evaluating the essays. Since the raters were from different cities around Iran, a traditional face-to-face classroom could not be held to train the raters for the rating methods used. Thus, the researcher sent the rating scale for the raters through email and then held a training session to inform them regarding the purpose of the study and the rater-training plan using Skype or WhatsApp for every rater. The training session lasted approximately 25 minutes for each rater.

To follow the previous research, a rater-training plan was developed according to the one presented by Barkaoui (2010b). In this training session, the raters were given enough information about the aim and the context of the study. Besides, all the details and issues related to each rating scale was expounded upon for the raters, zooming in on the descriptors and the writing tasks involved. Then, at the end of each session, some 30-minute time was allowed for the discussion and question and answer. In this part, all the participants talked about and discussed the rating scale regarding their expectations and writing tasks. Once the assessed writing samples were collected, the participants were asked to respond to the NEO-FFI inventory. This was done to gather information on the raters' personalities. The gathered data was analyzed utilizing MLM.

3.5. Data Analysis Procedure

In order to provide reliable answers to the research questions, the researcher decided to use multi regression modelling. This model was particularly used to overcome the deficiencies of the previous approaches such as analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), multiple regression analysis (MR), G-theory as well as Multi-Faceted Rasch model (MFRM).

As Barkaoui (2013) argues, there are some problems as well as limitations regarding these traditional approaches, the most prominent of which is the unit and level of analysis. He further explains that if we do not consider the levels, we will have to make a decision about the analysis unit and the degree to which the observations depend on each other. Using the multi regression or hierarchical structure of the data would help us get out of this dilemma.

Therefore, the nested data structure should be considered. For example, in this study the Level-1 (ratings) is nested in the Level-2 (raters) and the variables in different levels were related to them as well. The M-plus software was used to analyze the data gathered and manipulated through MLM approach.

4. Results and Discussion

Multi regression model (MRM) is proper for research cases in which data sources are categorized for participants exceeding one level (i.e., nested data). The analysis' units are commonly persons (at lower levels) who are placed within larger contextual/aggregate units (at higher levels). Further, individuals' measurements might be examined several times, while the lowest level of data in MR models is typically a person. Multi regression models, therefore, supply a different method of analysis for univariate or multi regression analysis of repeated measures. In growth curves, the main focus of the investigation is on the individual differences. Before testing treatment differences, multi regression models, in addition, can be used as an alternative to ANCOVA, where scores of the students on the measured variable are adapted for covariates (e.g. individual differences). These experiments can be analyzed by multi regression models without meeting the requirements of homogeneity-of-regression slopes which is needed when using ANCOVA. The initial model format is presented in Figure 1 for predicting the possible relationship between dependent (analytic-holistic scores) and independent variables (rater's professional, educational and personal background and gender) under the study.

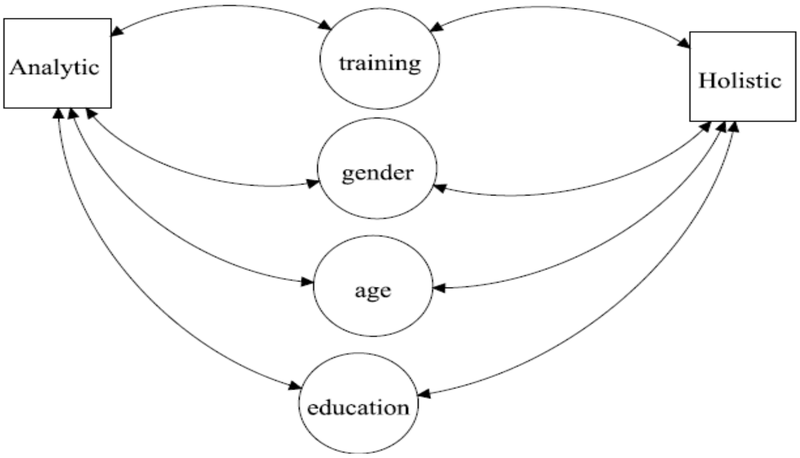

Figure 1: Initial Model Format

Research Question 1: To find out which background factors (personal background, professional background and work experience) at the rater level predict the score variability related to holistic method at the rating level, the following MRM analysis was run that will be responded by a number of statistical analyses. Table 2 displays descriptive statistics.

Table 2: Descriptive Statistics

Descriptive Statistics			
	Mean	Std. Deviation	N
Holistic Rating Score	2.9680	0.37318	80
Rater Age	42.0875	4.78974	80
Rater Professional Background	1.1875	0.39277	80
Rater Educational Background	1.8750	0.62389	80
Rater Training Background	1.5250	0.72871	80

According to Table 2, Holistic rating score M= 2.9 reflects below medium level of rating made by the raters and it is far below the exemplary rating score of 8 and SD=0.37 reflects the homogeneity of the group whose holistic score does not differ significantly in a group of 80 participants.

Table 3 confirms the predicted correlation between holistic score on the part of the raters and independent variables (rater professional background r = -0.012, educational background r =0.863, training background r = -0.0816) based on *Pearson Correlation* analysis. Accordingly, the correlation analysis confirms the significance of predicted correlation between holistic score on the part of the raters and independent variables (rater professional background p = 0.456, educational background p =0.000, training background p = 0.000) based on *1-tailed significance of Correlation*. Since the observed values between variables are less than 0.7, we can interpret that they are not multicollinear. Multicollinearity is a situation in which the relationship between the explanatory variables in a multiple regression model is very high. In simpler terms, perfect multicollinearity happens when, for instance as in the equation below, the correlation between two dependent variables is equal to 1 or −1. Additionally, we want the predictor variables to correlate with the outcome variable (in this case holistic rating score) at a value greater than 0.3, and Table 3 reveals that educational background *r* =0.863, training background *r* = -0.0816 have met those assumptions.

Table 3: Holistic Rating Score Correlations

Correlations

		Holistic Rating Score	Rater Age	Rater Professional Background	Rater Educational Background	Rater Training Background
Pearson Correlation	Holistic Rating Score	1.000	0.100	-0.012	0.863	-0.816
	Rater Age	0.100	1.000	0.381	0.021	-0.082
	Rater Professional Background	-0.012	0.381	1.000	-0.006	0.006
	Rater Educational Background	0.863	0.021	-0.006	1.000	-0.717
	Rater Training Background	-0.816	-0.082	0.006	-0.717	1.000
Sig. (1-tailed)	Holistic Rating Score	.	0.189	0.456	0.000	0.000
	Rater Age	0.189	.	0.000	0.428	0.234
	Rater Professional Background	0.456	0.000	.	0.477	0.481
	Rater Educational Background	0.000	0.428	0.477	.	0.000
	Rater Training Background	0.000	0.234	0.481	0.000	.
N	Holistic Rating Score	80	80	80	80	80
	Rater Age	80	80	80	80	80
	Rater Professional Background	80	80	80	80	80
	Rater Educational Background	80	80	80	80	80
	Rater Training In Background	80	80	80	80	80

According to Table 4 *adjusted R square* reflects a small sample size since the minimum sample size was 80. According to the *R square*, our model explains 0.828 of the variance of the dependent variable, hence significant *p* = 0.000.

Table 4: Holistic Rating Score Model Summary

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	0.910[a]	0.828	0.819	0.15868	0.828	90.484	4	75	0.000

a. Predictors: (Constant), Rater Training Background, Rater Professional Background, Rater Age, Rater Educational Background

b. Dependent Variable: Holistic Rating Score

Table 5 also displays a statistically significant *p* =0.000 value that confirms the predicted model at holistic score rating level and *F = 90.48 ratio* reveals mean square values that confirms the null hypothesis as it is close to 1.0. It's worth mentioning that a large *F* ratio implies that the variation existing among the means of group is beyond what we expect as occurring by chance.

Table 5: Multi Regression Model

Multi Regression Model						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9.113	4	2.278	90.484	0.000^b
	Residual	1.888	75	0.025		
	Total	11.002	79			

a. Dependent Variable: Holistic Rating Score

b. Predictors: (Constant), Rater Training in Background, Rater Professional Background, Rater Age, Rater Educational Background

Table 6 displays contributions of independent variables and we observe that rater' background education has the highest rate *t =8.38* and it has statistically significant relationship with holistic rating score.

Table 6: Coefficient Holistic rating score

Coefficients^a									
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	2.446	0.208		11.787	.000			
	Rater Age	0.005	0.004	.068	1.299	0.198	0.100	0.148	0.062
	Rater Professional Background	-0.031	0.049	-.032	-0.625	0.534	-0.012	-0.072	-0.030
	Rater Educational Background	0.345	0.041	0.576	8.386	0.000	.863	0.696	0.401
	Rater Training Background	-0.203	0.035	-0.397	-5.752	0.000	-.816	-0.553	-0.275

a. Dependent Variable: Holistic Rating Score

Table 7 displays standardized residual minimum = -3.07 and maximum = 2.14 that is within the assumptions and Cook's distance displays minimum = 0.000 and maximum =0 .129 where we do not expect a value more than 1 that reflects the value is ok here as well. If the Cook's vale was greater than 1 it could negatively violate the model under the analysis.

Table 7: Residuals Statistics Holistic rating score

Residuals Statistics					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.3357	3.4943	2.9680	0.33965	80
Std. Predicted Value	-1.862	1.550	0.000	1.000	80
Standard Error of Predicted Value	0.025	0.060	0.038	0.010	80
Adjusted Predicted Value	2.3235	3.4967	2.9678	0.33992	80
Residual	-0.48722	0.34039	0.00000	0.15461	80
Std. Residual	-3.070	2.145	0.000	0.974	80
Stud. Residual	-3.136	2.173	0.001	1.001	80
Deleted Residual	-.50824	.35236	0.00018	0.16344	80
Stud. Deleted Residual	-3.342	2.230	0.004	1.021	80
Mahal. Distance	0.958	10.173	3.950	2.493	80
Cook's Distance	0.000	0.103	0.011	0.022	80
Centered Leverage Value	0.012	0.129	0.050	0.032	80

a. Dependent Variable: Holistic Rating Score

Figure 2 displays the results of the study in terms of probability plot. The points more or less follow the line, except for a few deviations, they generally appear to follow the same line.

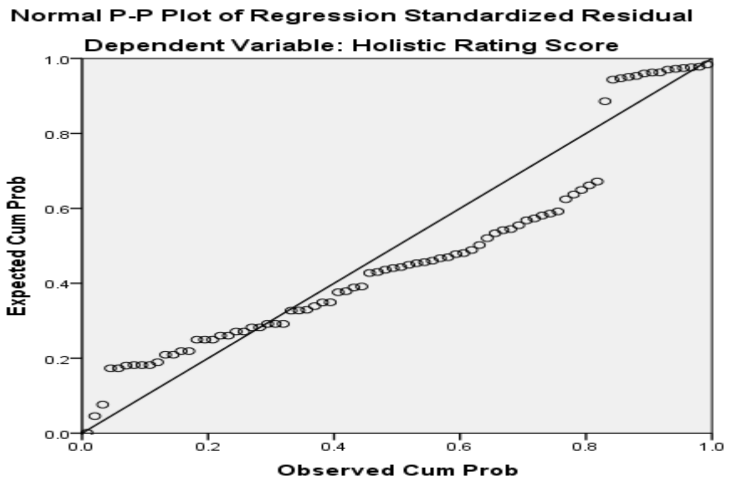

Figure 2: Normal P-P Plot of Regression Standardized Residual Holistic rating Score

In Figure 3, regression standardized residual is displayed on the *y* axis, while on the *x* axis the regression standardized predicted value is presented. As can be seen, none of the points fall out of the point -3 and 3 either on the *x* axis or the *y* axis. Thus, in this case, we are in a good shape and neither of the values are greater than 3, and neither of the values are less than -3.

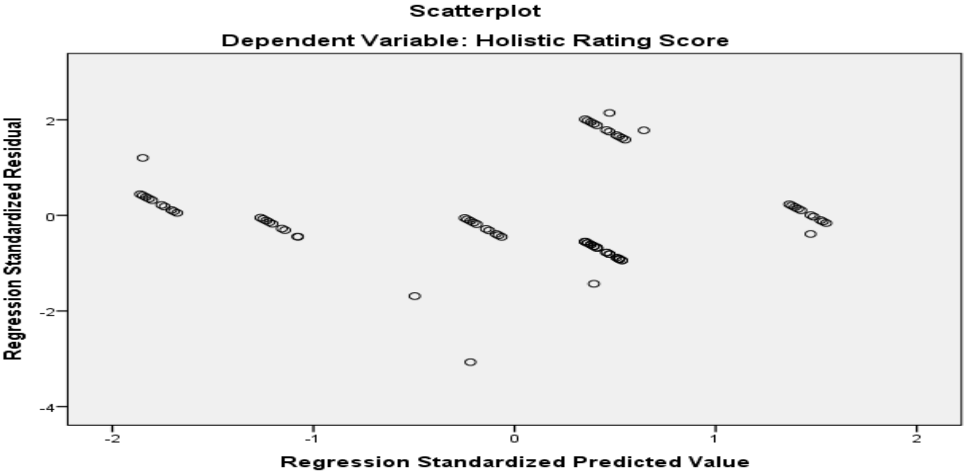

Figure 3: Scatterplot Holistic rating score

Exploring Question 2**:** In this part, MRM analysis was run, responded by a number of statistical analyses. The results of this analysis are reported in the Table 8.

Table 8: Descriptive Statistics

Descriptive Statistics			
	Mean	Std. Deviation	N
Analytic Rating Score	55.1566	3.13596	80
Rater Age	42.0875	4.78974	80
Rater Professional Background	1.1875	0.39277	80
Rater Educational Background	1.8750	0.62389	80
Rater Training In Background	1.5250	0.72871	80

According to Table 8, analytic rating score M= 55.15 reflects medium level of writing quality in rated articles scored by the raters, and it is far below the top rating score of 100 and SD=3.13, thus it reflects the heterogeneity of the group performance in terms of writing whose analytic score differ significantly in a group of 80 participants. Table 9 confirms the predicted correlation between analytic score on the part of the raters and independent variables (rater professional background *r* = -0.001, educational background *r* = -0.814, training background *r* = 0.0839) based on *Pearson Correlation* analysis.

Accordingly, the correlation analysis confirms the significance of predicted correlation between analytic score on the part of the raters and independent variables (rater professional background *p* = 0.496, educational background *p* =0.000, training background *p* =0.000) based on *1-tailed significance of Correlation*. Since the observed values between variables are less than 0.7 we can interpret that they are not multicollinear. We have perfect multicollinearity if, for example as in the equation below, the correlation between two dependent variables is equal to 1 or −1. Additionally, we want the predictor variables to correlate with the outcome variable (in this case analytic rating score) at a value greater than 0.3 and Table 8 reveals that educational background *r* = -0.814, training background *r* = 0.0839 have met those assumptions.

Table 9: Analytic Rating Score Correlations

Correlations						
		Analytic Rating Score	Rater Age	Rater Professional Background	Rater Educational Background	Rater Training In Background
Pearson Correlation	Analytic Rating Score	1.000	-0.064	-0.001	-0.814	0.839
	Rater Age	-0.064	1.000	0.381	0.021	-0.082
	Rater Professional Background	-0.001	0.381	1.000	-0.006	0.006
	Rater Educational Background	-0.814	0.021	-0.006	1.000	-.717
	Rater Training Background	0.839	-0.082	0.006	-0.717	1.000
Sig. (1-tailed)	Analytic Rating Score	.	0.285	0.496	0.000	0.000
	Rater Age	0.285	.	0.000	0.428	0.234
	Rater Professional Background	0.496	0.000	.	0.477	0.481
	Rater Educational Background	0.000	0.428	0.477	.	0.000
	Rater Training In Background	0.000	0.234	0.481	0.000	.
N	Analytic Rating Score	80	80	80	80	80
	Rater Age	80	80	80	80	80
	Rater Professional Background	80	80	80	80	80
	Rater Educational Background	80	80	80	80	80
	Rater Training In Background	80	80	80	80	80

Multi regression model analytic rating score is shown in Table 10. Also, coefficients analytic rating score is reported in Table 11.

Table 10: Multi Regression Model Analytic Rating Score

Multi Regression Modeling						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	619.230	4	154.808	73.637	0.000[b]
	Residual	157.673	75	2.102		
	Total	776.903	79			
a. Dependent Variable: Analytic Rating Score						
b. Predictors: (Constant), Rater Training Background, Rater Professional Background, Rater Age, Rater Educational Background						

Table 11: Coefficients Analytic Rating Score

Coefficients[a]									
	Unstandardized Coefficients		Standardized Coefficients			Correlations		
Model	B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
1 (Constant)	56.193	1.896		29.634	0.000			
Rater Age	-0.008	0.037	-0.011	-0.203	0.840	-0.064	-0.023	-0.011
Rater Professional Background	-0.019	0.450	-0.002	-0.042	0.967	-0.001	-0.005	-0.002
Rater Educational Background	-2.204	0.376	-0.439	-5.867	0.000	-0.814	-0.561	-0.305
Rater Training Background	2.253	0.323	0.524	6.978	0.000	0.839	0.627	0.363
a. Dependent Variable: Analytic Rating Score								

Table 11 displays contributions of independent variables and we observe that rater' background training has the highest rate *t* =6.97 and it has statistically significant relationship with holistic rating score. Residual statistics analytic rating score is performed in Table 12. As is seen in Table 12, standardized residual minimum = -2.73 and maximum = 3.38 that is within the assumptions and Cook's distance displays minimum = 0.000 and maximum = 0.141 where we do not expect a value more than 1 that reflects the value is proper here as well. If the Cook's vale was greater than 1, it could negatively violate the model under the analysis.

Table 12: Residual Statistics Analytic rating score

Residuals Statistics[a]					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	51.4114	60.4582	55.1566	2.79971	80
Std. Predicted Value	-1.338	1.894	.000	1.000	80
Standard Error of Predicted Value	.228	.545	.352	.089	80
Adjusted Predicted Value	51.4167	60.8023	55.1634	2.79245	80
Residual	-3.95824	4.91313	.00000	1.41275	80
Std. Residual	-2.730	3.389	.000	.974	80
Stud. Residual	-2.846	3.461	-.002	1.003	80
Deleted Residual	-4.30230	5.12504	-.00679	1.49640	80
Stud. Deleted Residual	-2.993	3.750	-.002	1.025	80
Mahal. Distance	.958	10.173	3.950	2.493	80
Cook's Distance	.000	.141	.012	.020	80
Centered Leverage Value	.012	.129	.050	.032	80
a. Dependent Variable: Analytic Rating Score					

In this part, the results of the study are presented in terms of probability plot as shown in Figure 4.

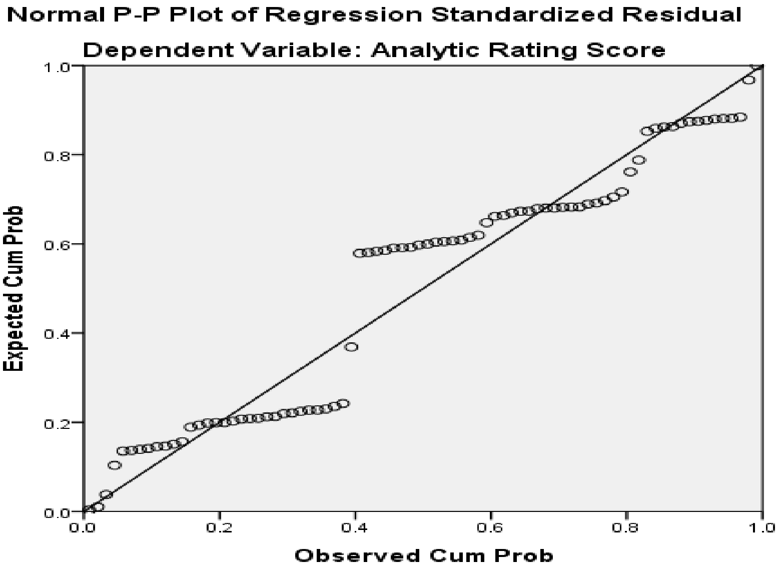

Figure 4: Normal P-P Plot of Regression Standardized Residual Analytic Rating Score

The points more or less follow the line, except for a few deviations, they generally appear to follow the same line. Furthermore, regression standardized residual was measured on the *y* axis, and the regression standardized predicted value was determined on the *x* axis, the results of which are shown in Figure 5. As is displayed in Figure 5, we see that none of the points fall out of the point -3 and 3 either on the *x* axis or the *y* axis. So in this case, we are in a good shape and neither of the values are greater than 3 and neither of the values are less than -3.

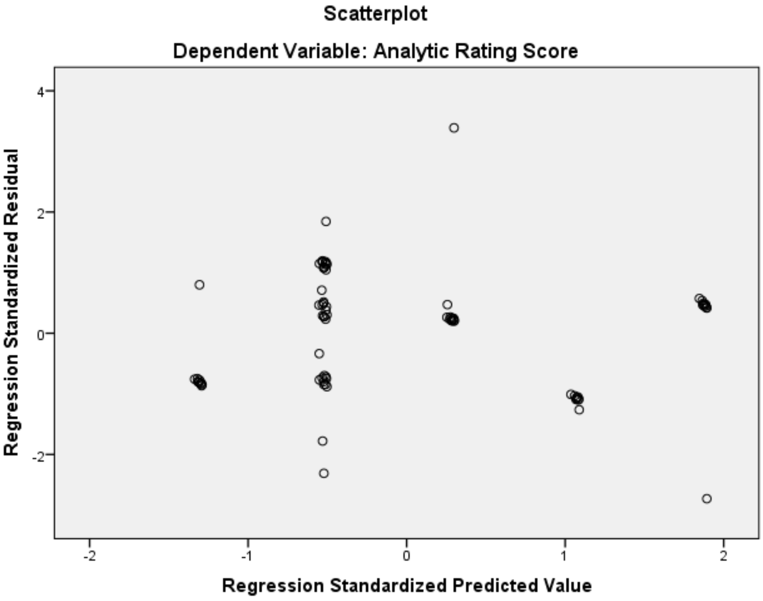

Figure 5: Scatterplot Analytic Rating Score

Figure 6 displays the final MRM format. As is displayed in Figure 6, we observe that there is a relationship as predicted by the study among the dependent variables (analytic and holistic rating score) and independent variables (professional background, educational background, personal background, and gender).

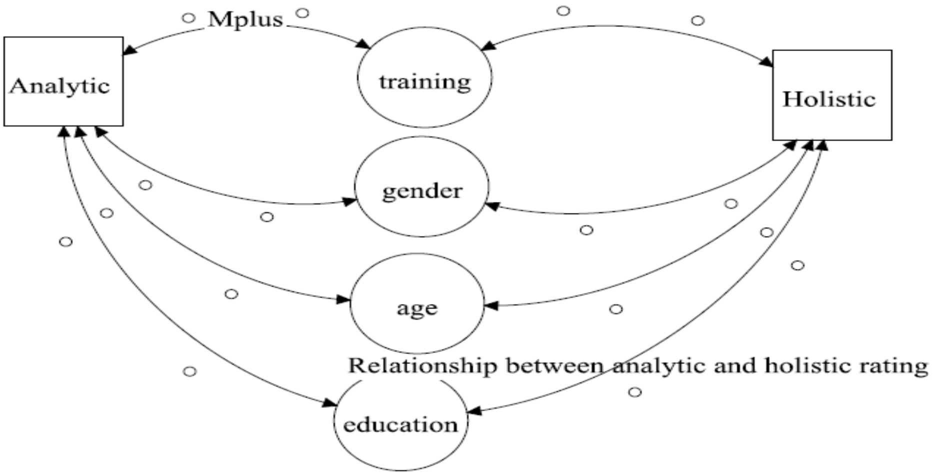

Figure 6: Final MRM Format Displaying Predicted Relationships

Research Question 3 and 4: According to Table 6 that shows the contributions of independent variables, it can be interpreted that the rater's background education has the highest rate *t =8.38* and it has a statistically significant relationship with holistic rating score. Therefore, the rater's background education predicts the score variability related to holistic method at the rating level.

According to Table 11, which shows contributions of independent variables, it can be concluded that the rater's background training has the highest rate *t* =6.97 and it has statistically significant relationship with holistic rating score.

Research Questions 5 and 6: According to the obtained results, there is a statistically significant relationship between personality factor and analytic/holistic rating scores, but these independent variables are less significant compared to educational and training background variables. As mentioned in response to research questions 3 and 5, background factors namely educational and training background of the raters significantly predict the raters scores both at analytic and holistic rating.

5. Conclusion and Implications

Today, English writing skill is becoming an integral part of the lives of literate people; thus, they are required to improve their writing skills in order to advance in technologically-based communications, such as chat rooms, emails, forums etc. The primary need of human to enjoy reading, other than these vehicles of digital communication, necessitates the counterpart, namely writing. Writing skill helps people live immortal. Therefore, applying efficient approaches to assess writing skills in EFL learners seems to be a necessity of scientific and educational communities in order to improve the quality of communication. So far, many studies have been conducted in the field of writing evaluation of EFL, which has examined this issue based on different aspects.

However, in this study, we have examined the issue of assessing writing skills for EFL learners from a different perspective, at which previous studies have not looked. In fact, this difference of opinion can indicate a new perspective in order to examine the skills of learners. In our view, the main purpose is to investigate variability impact in evaluating writing skills. For this purpose, we have investigated the role of personality types, background and scoring methods in writing rating of Iranian EFL students.

To achieve the goals of this study, six different questions have been asked. Then, by performing some tests and providing answers to each question, achievements are provided for this paper. In this study, we should say that the Big Five was used that is a well-known test. The results could possibly assist EFL writing raters and teachers in the language institutes well as university professors in the classroom context. It further provides implications for assessing EFL essays regarding the ways we should use holistic or analytic approaches in evaluating high-stakes writing projects and assist us in running rater training courses. According to the findings, the multiple-trait scoring method the potential to specify the level of students more justly, though both scoring methods measured the same construct.

Furthermore, empirical results indicate that a scale with more detailed levels of depicters offers the raters with significantly higher rater reliability than what happens in a scale with fewer particular depicters. On the other hand, these results show that the essays communicative quality was more significant than other writing's dimensions for both inexpert and experienced raters. Whereas the experienced raters were stricter considering grammatical accurateness, inexpert raters, moreover, were more easy-going than experienced raters in allocating more importance to argumentation.

This study like the other studies suffers from some limitations that examining each of them can be discussed as an open topic in the future. First, the participants of the study were selected from non-native raters and they were 50 male and female raters from Sistan and Baluchestan province. Second, in this study just the role of 3 factors namely, personality traits, scoring rubrics and scoring methods in writing rating were studied and other factors were not considered. Thirdly, finding the professional raters for the purpose of the study was very hard and it was another limitation of the study. Some of the raters for cooperating with researcher demanded some money for dedicating their time and best attempts, so the researcher invested some money for getting the reliable and valid results and findings. Another limitation of this study was the interest of the students on the essays' topics that might lead to their poorer performance in the case of uninterested topics. Lastly, this small-scale project was based on the data gathered from a small sample of teachers at a language center, as a result, not many generalizations can be made.

References

Ahmed A. (2018) Assessment of EFL Writing in Some Arab World University Contexts: Issues and Challenges. In Ahmed A., Abouabdelkader H. (eds) Assessing EFL Writing in the 21st Century Arab World. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-64104-1_1

Bachman, L. F., Palmer, A. S., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, *44*(1), 31-57. https://doi.org/10.5054/tq.2010.214047

Barkaoui, K. (2010b). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54-74. https://doi.org/10.1080/15434300903464418

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. *The companion to language assessment*, *3*, 1301-1322. https://doi.org/10.1002/9781118411360.wbcla070

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, *4*(1), 265.

Elahinia, H. (2004). *Assessment of writing through portfolios and achievement tests*. (Unpublished MA thesis), Teacher Training University, Tehran, Iran.

Fathi, J. & Khodabakhsh, M.R. (2019). The role of self-assessment and peer-assessment in improving writing performance of Iranian EFL students. *International Journal of English language and Translation Studies.7*(3), 1-10.

Ghanbari, N., & Barati, H. (2020). Development and validation of a rating scale for Iranian EFL academic writing assessment: A mixed-methods study. *Language Testing in Asia*, *10*(1), 1-21.

Glušac, T., & Milić, M. (2021). Quality of written instructions in teacher-made tests of English as a foreign language. *English Teaching and Learning*, https://doi.org/10.1007/s42321-021-00079-1

Han, T. & Ege, Ý. (2013). Using generalizability theory to examine classroom instructors' analytic evaluation of EFL writing. *International Journal of Education*, *5*(3), 20-35.

Howard, J.L., Gagné, M., Van den Broeck, A., Guay, F., Chatzisarantis, N., Ntoumanis, N. & Pelletier, L.G. (2020). A review and empirical comparison of motivation scoring methods: An application to self-determination theory. *Motivation and Emotion*, *44*(4), 534-548. https://doi.org/10.1007/s11031-020-09831-9

Huang, J. & Han, T. (2013). Holistic or analytic – A dilemma for professors to score EFL essays? *Leadership and Policy Quarterly*, *2*(1), 1-18.

Huang, J. (2010). Grading between lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly*, *7*(3), 219-233. https://doi.org/10.1080/15434300903540894

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, *17*(3), 123-139. https://doi.org/10.1016/j.asw.2011.12.003

Khodabandeh, F. (2021). The comparison of mind mapping-based flipped learning approach on introvert and extrovert EFL learners' speaking skill. *Iranian Journal of English for Academic Purposes*, *10*(1), 35-53.

Knoch, U., Read, J., & Randow, J. V. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, *20*, 148-161. https://doi.org/10.1016/j.jslw.2011.02.001

Li, H., Xiong, Y., Hunter, C.V., Guo, X. & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education*, *45*(2), 193-211. https://doi.org/10.1080/02602938.2019.1620679

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543-560.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Peter Lang, Frankfurt (2005) (Volume 3, Language Testing and Evaluation Series, ... A Book Review on Learner Identity and Beliefs in EFL Writing. *Journal of English Learner Education*, *11*(1), 104-109. https://orcid.org/0000-0002-3096-2046

Peker, H. & Torlak MA, M. (2020). A Book Review on Learner Identity and Beliefs in EFL Writing. *Journal of English Learner Education*, *11*(1), 104-109. https://orcid.org/0000-0002-3096-2046

Pierce, L. V., & O'Malley, J. M. (1992). *Performance and portfolio assessment for language minority students* (Vol. 9). National Clearinghouse for Bilingual Education.

Ragupathi, K., & Lee, A. (2020). Beyond fairness and consistency in grading: The role of rubrics in higher education. Edited by: Catherine Shea Sanger, Nancy W Gleason. In *Diversity and inclusion in global higher education* (pp. 73-95). Palgrave Macmillan, Singapore. https://doi.org/10.1007/978-981-15-1628-3

Rakedzon, T., & Baram-Tsabari, A. (2017). To make a long story short: A rubric for assessing graduate students' academic and popular science writing skills. *Assessing Writing*, *32*, 28-42.

https://doi.org/10.1016/j.asw.2016.12.004

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language testing*, *22*(1), 1-30.

Shabani, E. A., & Panahi, J. (2020). Examining consistency among different rubrics for assessing writing. *Language Testing in Asia*, *10*(1), 1-25. https://doi.org/10.1186/s40468-020-00111-4

Siyyari, M. (2011). The big five personality traits: A new horizon of research in language teaching. *Iranian EFL Journal*, *7*(6), 283-295.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the development of children*, *23*(3), 34-41.

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

Yeung, A., Read, J., & Schmid, S. (2012, November). Students' learning styles and academic performance in first year chemistry. In *Proceedings of the Australian Conference on Science and Mathematics Education (formerly UniServe Science Conference)*.

Zhang, B., Xiao, Y., & Luo, J. (2015). Rater reliability and score discrepancy under holistic and analytic scoring of second language writing. *Language Testing in Asia*, *5*(1), 1-9. https://doi.org/10.1186/s40468-015-0014-4

Zhou, S. N., Liu, Q. Y., Koenig, K., Li, Q., & Bao, L. (2021). Analysis of two-tier question scoring methods: a case study on the Lawson's classroom test of scientific reasoning. *Journal of Baltic Science Education*, *20*(1), 146-159. http://oaji.net/articles/2021/987-1611649216.pdf

Appendix A: Essay Writing Topic

A person's worth nowadays seems to be judged according to social status and material possessions. Old-fashioned values, such as honor, kindness and trust, no longer seem important.

To what extent do you agree or disagree with this opinion?

Give reasons for your answer and include any relevant examples from your own knowledge or experience.

Appendix B: Background Questionnaire

Personal background
Please provide your Personal background in this section.
Gender: Age: Field of study: Degree:
Professional training in rating
Please answer the following questions.
Do you pass any professional training for rating writing tasks?
How many years do you rate IELTS writing or other writings?
What type of scoring method do you apply for rating? Analytic or holistic? And why?
Do you have any personal scoring method for rating?
Work experience
Please answer the following questions.
How many years do you teach English?
How many years do you teach IELTS?
What English courses do you teach?
Where do you teach English language?
Do you teach English courses in university?