Porta Linguarum 38, June 2022 27-45

Using generalizability theory to investigate the variability and reliability of EFL composition scores by human raters and e-rater

Elif Sari
ORCID: 0000-0002-3597-7212
Karadeniz Technical University
Turgay Han
Ordu University

Received: 22 January 2021 / Accepted: 20 June 2022
DOI: 10.30827/portalin.vi38.18056
ISSN paper edition: 1697-7467, ISSN digital edition: 2695-8244

ABSTRACT: Using the generalizability theory (G-theory) as a theoretical framework, this study aimed at investigating the variability and reliability of holistic scores assigned by human raters and e-rater to the same EFL essays. Eighty argumentative essays written on two different topics by tertiary level Turkish EFL students were scored holistically by e-rater and eight human raters who received a detailed rater training. The results showed that e-rater and human raters assigned significantly different holistic scores to the same EFL essays. G-theory analyses revealed that human raters assigned considerably inconsistent scores to the same EFL essays although they were given a detailed rater training and more reliable ratings were attained when e-rater was integrated in the scoring procedure. Some implications are given for EFL writing assessment practices.
Key words: EFL writing assessment, generalizability theory, scoring variability, scoring reliability, automated writing evaluation (AWE).

Uso de la teoría de la generalización para investigar la variabilidad y confiabilidad de las puntuaciones de composición de EFL por evaluadores humanos y e-rater

RESUMEN: Utilizando la teoría de la generalización (teoría G) como marco teórico, este estudio tuvo como objetivo investigar la variabilidad y confiabilidad de los puntajes holísticos asignados por evaluadores humanos y e-rater a los mismos ensayos de inglés como lengua extranjera. Ochenta ensayos argumentativos escritos sobre dos temas diferentes por estudiantes turcos de inglés como lengua extranjera de nivel terciario fueron calificados de manera integral por un evaluador electrónico y ocho evaluadores humanos que recibieron una capacitación detallada como evaluador. Los resultados mostraron que los evaluadores electrónicos y humanos asignaron puntajes holísticos significativamente diferentes a los mismos ensayos de inglés como lengua extranjera. Los análisis de la teoría G revelaron que los evaluadores humanos asignaron puntajes considerablemente inconsistentes a los mismos ensayos de inglés como lengua

extranjera, aunque se les proporcionó una capacitación detallada para los evaluadores y se obtuvieron calificaciones más confiables cuando el evaluador electrónico se integró en el procedimiento de puntaje. Se dan algunas implicaciones para las prácticas de evaluación de escritura EFL.
Palabras clave: evaluación de redacción de inglés como lengua extranjera, teoría de la generalización, variabilidad de puntuación, fiabilidad de puntuación, evaluación de escritura automatizada.

1. Introduction

Reliability and validity are two concepts that most influence the quality of an assessment procedure (Hyland, 2003). Reliability is the consistency of test takers' scores when they are tested on different occasions, evaluated through different tasks, or scored by different raters (Johnson, Penny, & Gordon, 2009). Validity refers to the accuracy of interpretations made based on the test scores (Bachman, 1990). Although getting consistent scores from a test does not ensure that the test measures what it asserts to measure, reliability is a prerequisite for validity (Popham, 1981). Therefore, scoring reliability should be regarded "as a cornerstone of sound performance assessment" (Huang, 2008, p. 202).

Regarding assessing writing performance, the research has reported it to be a difficult task because of several factors that contribute to the error score such as the different aspects of writing performance (e.g., social context), the rubric type (e.g., holistic or analytic), and the language proficiency, conceptual knowledge, and judgemental ability of the students. Additionally, raters can have different rating behaviours, different decision-making processes, and different scoring tendencies (e.g., tendency to give lower or higher scores) (e.g., Baker, 2010; Heaton, 2003; Han, 2013; Lim, 2009). Raters' L1 and previous rater training are also among the factors that affect scores assigned to a piece of writing (Chang, 2002; Shi, 2001). Thus, one of the common source for error score can be rater subjectivity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014). Different raters may assign different scores to the same essay (i.e., inter-rater reliability), or the same rater may assign different scores to the essays which are of the same quality (i.e., intra-rater reliability), which will decrease the reliability of the scores (Brown, 2004; Homburg, 1984).

A higher degree of reliability should be ensured when the test scores are used to make high-stakes decisions that are not easily reversed (AERA, APA, & NCME, 2014). In order to increase the reliability of scores, two or more raters are suggested to be involved in the writing assessment procedure after they receive rater training to interpret a specific rating scale in a consistent way (Blood, 2011). In addition, training and feedback sessions should be repeated at certain intervals (Weigle, 2002). However, this is difficult to apply in most situations as it is not time-efficient and cost-effective (Attali & Burstein, 2006). In addition, raters may have some unconscious biases that are resistant to be corrected through training (Blood, 2011). In this sense, Automated Essay Scoring (AES) systems have been designed in an attempt to provide an instant, cost-effective, and reliable writing assessment (Chodorow

& Burstein, 2004; Latifi & Gierl, 2020). Various automated scoring systems (e.g., Intelligent Essay Assessor by Pearson Knowledge Technologies, e-rater by Educational Testing Service, and Intellimetric by Vantage Learning) (Warschauer & Ware, 2006) are being used to evaluate thousands of essays in both high-stakes standardized tests (e.g., TOEFL, GRE, or TWE) and low-stakes classroom assessment for educational purposes (Hockly, 2019; Shermis et al., 2010). All of these AES systems were trained on essays scored by human raters to extract the features which predict human scoring so that they could measure these features to make prediction on a new essay (Chodorow & Burstein, 2004).

The AES system used in the current study is "e-rater" that was developed by English Testing Service (ETS). E-rater was first launched in 1999 and used as one of the two raters in the writing section of the Graduate Management Admissions Test (GMAT). Recently, e-rater has been used as a co-rater in the L2 writing sections of large-scale standardized tests such as the Test of English as a Foreign Language (TOEFL) and the Graduate Record Examination (GRE) in the same way as it was used in the GMAT (Bridgeman et al., 2012). Also, a version of e-rater called the Criterion Online Writing Evaluation Service was created by ETS in order to help students plan, write, and revise their essays in writing instruction. E-rater uses "Natural Language Processing" to extract some features of writing (e.g., syntactic variety and the organization of ideas) from the essays which were pre-scored by expert human raters. The computer system conducts regression analysis to determine the best combinations of these features to imitate the scores assigned by expert human raters; then, these combinations are coded into the computer program to assess new essays (Attali & Burstein, 2006; Chodorow & Burstein, 2004).

The studies that focused on the reliability of scores obtained from e-rater were frequently conducted in high-stakes testing contexts where a large number of essays need to be scored in a short time and the reliability of scores is of great importance (Bridgeman et al., 2012). These studies were based on the agreement rates between human rater scores and e-rater scores since they regarded human scoring as "the gold standard" (Bridgeman et al., 2012, p.39). While some of these studies showed a considerable consistency between the scores given by human raters and e-rater (e.g., Burstein et al., 1998; Elliot, 2001; Foltz et al., 1999; Shermis et al., 2002), some other studies found conflicting results (e.g., Ebyary & Windeat, 2010; Hoang & Kunnan, 2016; Huang, 2014; James, 2006; Li et al., 2014; Liu & Kunnan, 2016).

Considering the previous research literature, although several studies have investigated the comparison between the scores assigned by human raters and e-raters, to our best knowledge, no study has investigated the score reliability and variations from different scoring scenarios that can be obtained in the different human rater-e-rater combinations (e.g., two human raters-e-rater, three human raters-e-rater, etc.) through using a more sophisticated model "Generalizability Theory" (G-theory) approach rather than the "Classical Test Theory (CTT)" approach. Thus, this study aims to bridge this research gap through investigating the sources of score variation as well as the reliability of scores when e-rater scores were integrated with different number of human rater scores in an EFL writing assessment context employing G-theory as a theoretical framework.

Briefly, using the G-theory as a framework for analysis, the purpose of this study was to examine how e-rater scores impacted the variability and reliability of holistic EFL scores when they were integrated with the scores assigned by different number of human raters in an EFL writing assessment context in a Turkish state university. Specifically, the following three research questions were asked in this study:

1. Are there significant differences between the holistic scores assigned by each of the eight human raters and e-rater to the same EFL essays?
2. Does the integration of e-rater scores with human rater scores impact the sources of score variation contributing to the holistic scores?
3. Does the reliability (e.g., generalizability coefficients for norm-referenced score interpretations) of the holistic scores differ when e-rater scores are integrated with different number of human rater scores?

1.1. Theoretical framework

The three theoretical frameworks that are used in ESL/EFL writing assessment research are The CTT approach, the item response theory (IRT approach), and the G-theory approach (Elorbany & Huang, 2012). The theoretical framework used for the current study is the G-theory approach. CTT estimates only two sources of errors, "a single ability and a single source of errors" (Bachman, 1990, p.188), and thus it is regarded as a weak theory (Huang, 2012). On the other hand, G-theory that was developed by Cronbach et al. (1972) with the aim of overcoming the limitations of CTT, can explain multiple sources of variation or error in measurement through a single analysis (Briesch et al., 2014; Shavelson & Webb, 1991). G-theory is a statistical analysis that enables to investigate the impact of each source of error and the interaction of multiple sources of error on the generalizability of scores obtained from an assessment (Shavelson & Webb, 1991). For instance, when two or more raters are required to score a number of essays written on two or more topics using two different rating methods (analytic and holistic), a set of sources cause the variability of their scores such as writer, rater, topic, scoring method, and the interaction between these sources. The sources that cause variability in scores are called facets and the levels of each source are identified as conditions in G-theory. For example, if rater is a facet, first rater, second rater, third rater, etc. are accepted as conditions (Briesch et al., 2014; Güler, Uyanık, & Teker, 2012).

The quantity of the variance arising from each facet can be measured through G-theory. This measurement consists of two phases: a generalizability study (G-study) and a decision study (D-study). D-study uses the estimates found in the G-study to develop more efficient measurement procedures for practical use (Kieffer, 1998). In summary, while the G-study examines the role of different sources of error in measurement and the impacts of some possible changes in the design of measurement, D-study provides the integration of the ideal design and interprets the score reliability (Briesch et al., 2014; Huang, 2008).

2. Methodology

This study used a quantitative research method to examine the variability and reliability of scores assigned by human raters and e-rater using the G-theory approach.

2.1. The selection of writing samples

The study used departmental writing test data provided by first-year English major students at the English Language Teaching Department of a state university in Turkey. The writing samples were obtained as follows. First, the students were informed that their essays would be used for the purpose of this study and their written consent was received. Official permission was also obtained from the Dean's Office of the Faculty where the students were enrolled. Then, the course teacher asked the students to respond to two different argumentative writing tasks in two different sessions (i.e., each student wrote one argumentative essay in each of the two sessions). In each session, the students were required to write a 300-to-350-word essay on a single topic that had been selected for all students. Both tasks were selected from the essay topics of "Criterion Topic Library" by the writing course teacher considering the students' level of proficiency, experiences of writing classes, educational interests, and cultural characteristics. The tasks were assumed to be parallel regarding the topic familiarity as they did not require background information. The selected essay topics had not been discussed with the students beforehand. The students were given 60 minutes to write their essays using Microsoft Word. The course teacher accepted the essays through a text-matching software, Turnitin, to ensure the originality of the essays. In total, 150 argumentative essays were collected from 75 students.

Second, with the purpose of maximizing the differences among the papers, two independent raters carefully divided the essays into three levels of quality (e.g., high, medium, and low) based on the 6-point holistic scoring scale provided by Criterion. Only the essays which both of the two raters grouped as high-quality or low-quality were selected. Finally, 40 high-quality and 40 low-quality essays written on the following topics were randomly selected by the researchers.

Topic 1

"What makes a professor great? Prominence in his or her field? A hot new book? Good student reviews every semester? What standards should be used to assess the quality of college faculty members? Support your position with reasons and examples from your own experiences, observations or reading."

Topic 2

"After they complete their university studies, some students live in their hometowns. Others live in different towns or cities. Which do you think is better — living in your hometown or living in a different town or city? Give reasons for your answer."

2.2. The selection of raters

A total of eight raters, six female and two male, participated in this study according to their voluntariness to take part in the study and their proximity to the researcher. They were all full-time employees at the school of Foreign Languages at a Turkish state university and had a bachelor's or master's degree in English Language Teaching. All of the raters were native speakers of Turkish and got a score of 90 and above from the Foreign Language Proficiency Exam called YDS, which is officially accepted as a national language proficiency test in Turkey. Their ages ranged between 31 and 45, with five between 31 and 35. They had at least five years of experience in writing instruction and assessment, but none of them had received a formal rater training prior to this study. All of the raters were informed

about the purpose of the study and they wholeheartedly agreed to participate in the study as they thought the results of the study would contribute to the writing assessment practices in higher education.

2.3. The rating scale

In scoring the essays, the 1–6-point holistic scoring scale (e.g., 1 is assigned to very poor-quality writing, 6 is assigned to the highest quality writing) created by ETS and used by e-rater to score the submitted essays was used by the human raters.

2.4. The rater training and rating procedure

Rater training, accompanied by a scoring rubric that specifies the criteria to be consulted while making judgements in assessing writing, is crucial to increase reliability of scores assigned to writing samples (Homburg, 1984; Weigle, 1994). All the participant raters were subjected to a rater training session, which lasted approximately two hours, before the scoring procedure started. The training session was conducted by the first author of the study, who had more than ten years of experience in teaching and assessing EFL writing and attended in-service rater training beforehand. A traditional rater training classroom model was followed as it is practical for small-scale writing assessment practices in research studies (Johnson et al., 2009). At the beginning of the training session, the raters were briefly informed about the purpose of the study and were given a consent form which assured their rights and the confidentiality of their identities, after which the holistic rating scale was reviewed. The elements of each level within the scale and what those elements meant, was discussed until the expectations were clear to each of the raters. After this discussion, a small-scale pilot study was conducted with three essays of different levels of quality (i.e., good, average, poor) to see whether the raters understood the holistic scoring scale. After the raters completed the scoring of the three essays, they discussed their scores and solved the disagreements if they had any. Then, the raters independently scored nine argumentative essays of different qualities, which were not included in the main data collection procedure. Finally, the raters were given data packs which included 80 argumentative essays that were printed on paper, one holistic scoring rubric, and a background information questionnaire. The raters were required to score the essays at their homes or offices in two sessions in the same day within two weeks. The rating procedure was conducted in the summer holiday period in order to alleviate rater fatigue due to their regular jobs.

2.5. Data analysis

2.5.1. Descriptive and inferential statistics

Descriptive statistical analysis (the mean and standard deviation) and paired sample t-tests were conducted for the holistic scores assigned by e-rater and each of the eight human raters on the same EFL essays. These analyses were carried out with the aim of examining if there were any significant mean score differences between e-rater scores and human rater scores.

2.5.2. G-theory analysis

Through the use of EduG computer program, this study used G-theory framework to examine the role of students, raters, and their interaction in the variance of the scores given by e-rater and human raters. In the current study, students were the object of measurement whereas raters were random facets. For all of the participating students (persons as p) created the essays and the same raters (r) scored all of the essays, the design of the G-study was fully crossed as (p x r). Person-by-rater (p x r) random effect G-study was carried out to acquire variance component estimates for independent sources of variation such as persons (p), raters (r), person-by-rater (p x r) for 80 essays evaluated through holistic scoring method. Furthermore, generalizability coefficient (used in norm-referenced tests) and dependability coefficient (used in criterion-referenced tests) calculations were administered for human rater scores and different combinations of human rater+e-rater scores to reveal whether score reliability shows a difference when e-rater is used together with human raters in the scoring procedure.

3. Results

3.1. Descriptive and inferential statistical results

Each of the papers was scored by e-rater and by each independent human raters using the same 6-point holistic scoring scale. Table 1 presents the descriptive statistics for the holistic scores given by e-rater and each human rater (HR) for Topic One and Topic Two.

Table 1. Descriptive statistics of the holistic scores given by e-rater and each human rater

		TOPIC ONE		TOPIC TWO	
Rater	N	Mean	SD	Mean	SD
E-rater	40	3.02	.91	5.65	.36
HR1	40	4.37	.77	4.57	.84
HR2	40	3.02	1.12	3.37	.83
HR3	40	2.90	1.17	2.22	.97
HR4	40	3.90	1.25	4.30	1.01
HR5	40	4.05	1.33	4.22	1.14
HR6	40	3.30	.93	3.52	.93
HR7	40	4.42	1.17	4.97	.89
HR8	40	3.37	1.37	4.92	.88

Table 1 shows that one human rater (HR3) gave lower holistic scores than e-rater and one human rater (HR2) assigned nearly the same holistic scores with those assigned by e-rater.

The other six human raters assigned higher scores than e-rater. However, for Topic Two, all of the human raters assigned lower holistic scores to essays than e-rater. Further, the standard deviation for Topic One is over 1 point for nearly all human raters except for two human raters, indicating that these human raters scored the papers very differently. Conversely, the standard deviation for Topic Two is lower than 1 point for nearly all human raters, except for two human raters, indicating that these human raters scored the papers consistently. Briefly, although nearly all human raters scored the Topic One papers higher than e-rater, the human raters scored less consistently, the reverse is true for the Topic Two papers.

In order to reveal whether there were any significant mean score differences between the holistic scores given by e-rater and each of the eight human raters to the same EFL essays on two topics, paired sample t-tests were conducted. Table 2 shows the t-test results for Topic One.

Table 2. Paired samples t-tests results for Topic One

	MEAN	SD	SE MEAN	t	df	p
Pair 1: E-rater-HR1	-1.3500	.8335	.1318	-10.24	39	.000*
Pair 2: E-rater-HR2	.0000	1.0127	.1601	.00	39	1.000
Pair 3: E-rater-HR3	.1250	1.3994	.2212	.56	39	.575
Pair 4: E-rater-HR4	-.8750	.8223	.1300	-6.72	39	.000*
Pair 5: E-rater-HR5	-1.0250	.9996	.1580	-6.48	39	.000*
Pair 6: E-rater-HR6	-.2750	.7840	.1239	-2.21	39	.032*
Pair 7: E-rater-HR7	-1.4000	1.0076	.1593	-8.78	39	.000*
Pair 8: E-rater-HR8	-.3500	.9486	.1500	-2.33	39	.025*

**Note: indicates significant difference at the significance level of .05.*

As can be seen in Table 2, there was no significant mean score difference between the holistic scores assigned by e-rater and the HR2 and HR3. For all other pairs, there existed a significant difference between the holistic scores given by e-rater and human raters. Further, in these pairs, human raters assigned significantly higher scores than e-rater. The greatest mean score difference was between e-rater and HR7 (-1.40). Table 3 shows the t-test results for Topic Two.

Table 3. Paired samples t-tests results for Topic Two

	MEAN	SD	SE MEAN	t	df	p
Pair 1: E-rater-HR1	1.2750	.9054	.1431	8.90	39	.000*
Pair 2: E-rater-HR2	2.4750	.8469	.1339	18.48	39	.000*
Pair 3: E-rater-HR3	3.6250	1.0048	.1588	22.81	39	.000*
Pair 4: E-rater-HR4	1.5500	.9323	.1474	10.51	39	.000*
Pair 5: E-rater-HR5	1.6250	1.1477	.1814	8.95	39	.000*
Pair 6: E-rater-HR6	2.3250	.8883	.1404	16.55	39	.000*
Pair 7: E-rater-HR7	.8750	.8529	.1348	6.48	39	.000*
Pair 8: E-rater-HR8	.9250	.7641	.1208	7.65	39	.000*

**Note: indicates significant difference at the significance level of .01.*

As can be seen in Table 3, there was a statistically significant difference between the holistic scores assigned by e-rater and each human rater. The greatest mean score difference was between e-rater and HR3. Contrary to the results obtained from the Topic One, e-rater assigned higher scores than human raters in all pairs. It can be concluded that human raters and e-rater assigned significantly different scores on the same essays on both Topic One and Topic Two, and topic has an important impact on the holistic scores assigned by e-rater and human raters.

3.2. The results of the person-by-rater random effects G-studies

Two different person-by-rater (p x r) random effects G-studies were carried out for the holistic scores assigned by human raters and a combination of human raters+e-rater to the same EFL essays. The purpose of these G-studies was to investigate the extent to which students and raters and the interactions between students and raters contribute to the variance of the holistic scores given by human raters and whether the scores given by e-rater impact the contribution of these variance components to the score variability. Table 4 and Table 5 show the results of G-studies for Topic One and Topic Two respectively.

Table 4. Variance components for random effects P X R design (Topic One)

RATER	VARIANCE SOURCE	df	σ2	%
Human Raters	P	39	0.66	39.7
	R	7	0.33	20.2
	PR	273	0.67	40.1
	Total	319		100
Human Raters	P	39	0.65	38.9
+	R	8	0.34	21.0
	PR	312	0.63	40.1
E-rater	Total	359		100

As indicated in Table 4, for the scores given by human raters, the residual was found to be the greatest variance component (40.1 % of the total variance), which indicates that a large unexplained source of variance existed in this design because of the interaction between raters and papers, and other systematic and unsystematic error sources. The second largest variance component was person (39.7 % of the total variance), implying that the students differed considerably in their writing abilities. The third largest component was found to be rater (20.2 % of the total variance), suggesting that the raters assigned markedly inconsistent scores to the essays. In addition, Table 4 shows that when the e-rater scores were mixed with human rater scores, the contribution of the variance components to the total variance was similar to that obtained for human rater scores on Topic One.

Table 5. Variance components for random effects P X R design (Topic Two)

RATER	VARIANCE SOURCE	df	σ2	%
Human Raters	P	39	0.29	17.2
	R	7	0.84	48.7
	PR	273	0.59	34.1
	Total	319		100
Human Raters + E-rater	P	39	0.25	13.1
	R	8	1.11	57.9
	PR	312	0.55	29.0
	Total	359		100

As shown in Table 5, for the scores given by human raters, rater was found to be the largest variance component (48.7 % of the total variance), indicating that the raters assigned significantly inconsistent scores to the essays. The second largest variance component was the residual (34.1 % of the total variance), which indicates that a large unexplained source of variance existed in this design because of the interaction between raters and papers, and other systematic and unsystematic error sources. The third largest component was found to be person (17.2 % of the total variance), suggesting that the students differed in their writing abilities. Moreover, when the e-rater scores were mixed with human rater scores, the contribution of the variance components to the total variance was similar to that obtained for human rater score on Topic Two.

3.3. Generalizability and dependability coefficients for human raters and e-rater

Using the person-by-rater (p x r) random effects design, the generalizability coefficients (symbolized as Ep2) and dependability coefficients (symbolized as ϕ) were calculated for each topic. Based on the idea that e-rater can replace one of the human raters to make the scoring process more cost-effective (Bridgeman et al., 2012), the generalizability coeffi-

cients and dependability coefficients were calculated for human rater scores and different combinations of human rater+e-rater scores separately with the purpose of seeing the impact of e-rater on score reliability. As Williamson et al., (2012) established, the coefficients that were at .70 or above were considered to refer to a high and meaningful reliability. Table 6 and 7 present the results for Topic One and Topic Two respectively.

Table 6. Generalizability and dependability coefficients for Topic One

NUMBER OF PAPERS	NUMBER OF RATERS	Ep2	%
40	2 HR	.65	.43
40	2 HR + E-rater	.70	.58
40	3 HR	.70	.48
40	3 HR + E-rater	.77	.58
40	4 HR	.78	.67
40	4 HR + E-rater	.81	.72
40	5 HR	.79	.71
40	5 HR + E-rater	.83	.76
40	6 HR	.82	.76
40	6 HR + E-rater	.86	.80
40	7 HR	.87	.81
40	7 HR + E-rater	.89	.83
40	8 HR	.89	.84
40	8 HR + E-rater	.90	.86

As seen in Table 6, the generalizability and dependability coefficients obtained for the current 40 essays and 8 human raters scenario (Ep2 = .89 and Φ = .84) were slightly lower than those obtained when e-rater scores were integrated (Ep2 = .90 and Φ = .86). It is also seen that the generalizability and dependability coefficients increased when e-rater scores were integrated with the human rater scores. While an acceptable level of generalizability coefficient (i.e., .70 and above) was achieved when e-rater scores were integrated with two human raters' scores, an acceptable level of dependability coefficient was achieved when e-rater scores were integrated with four human raters' scores.

Table 7. Generalizability and dependability coefficients for Topic Two

NUMBER OF PAPERS	NUMBER OF RATERS	Ep2	Φ
40	2 HR	.52	.52
40	2 HR + E-rater	.58	.46
40	3 HR	.68	.59
40	3 HR + E-rater	.70	.54
40	4 HR	.71	.68
40	4 HR + E-rater	.72	.64
40	5 HR	.74	.53
40	5 HR + E-rater	.74	.48
40	6 HR	.76	.51
40	6 HR + E-rater	.76	.46
40	7 HR	.78	.61
40	7 HR + E-rater	.78	.56
40	8 HR	.80	.62
40	8 HR + E-rater	.80	.58

As shown in Table 7, the generalizability coefficient obtained for the current 40 essays and 8 human raters scenario (Ep2 = .80) was the same as the generalizability coefficient which was obtained when e-rater scores were integrated (Ep2 = .80). However, dependability coefficient obtained for the current 40 essays and 8 human raters scenario (Φ = .62) was higher than that obtained when e-rater scores were integrated (Φ = .58). It is also seen that the dependability coefficients decreased when e-rater scores were integrated with human raters' scores although the generalizability coefficients mostly remained the same. This result was due to the high rater variability for Topic Two as it was revealed in Table 5 which shows the variance components for random effects P X R design. While an acceptable level of generalizability coefficient (i.e., .70 and above) was achieved when e-rater scores were integrated with three human raters' scores, an acceptable level of dependability coefficient was not achieved with the current scenario. Therefore, a D-study based on crossed design (i.e., p x r) was conducted for Topic Two in order to examine dependability coefficient for different rater scenarios. The number of raters increased until an acceptable level of dependability coefficient was achieved since it is estimated that when the number of facets is increased in a G-study design, higher generalizability and dependability coefficients will be acquired. Table 8 presents the generalizability and dependability coefficients when the number of raters were increased in different scenarios.

Table 8. Generalizability and dependability coefficients for different numbers of raters

	HUMAN RATERS		HUMAN RATERS + E-RATER	
Number of raters	Ep2	□	Ep2	□
10	.83	.68	.82	.60
12	.86	.71	.84	.64
14	.88	.74	.86	.68
16	.89	.77	.88	.71

According to Table 8, within the current G-study design, an acceptable level of dependability coefficient was achieved with the twelve-human rater scenario. When the e-rater scores were integrated with human rater scores, the dependability coefficient reached an acceptable level with the sixteen-human rater + e-rater scenario. This result shows that the scoring variability due to the rater is very high for Topic Two.

4. Discussion

The first research question aimed at investigating whether there were any differences between the holistic scores given by e-rater and each human rater to the same EFL essays. Paired samples t-tests showed that each human rater and e-rater assigned significantly different holistic scores to the same EFL essays. This result is contrary to the findings of previous research (Burstein et al., 1998; Elliot, 2001; Foltz et al., 1999; Shermis et al., 2002) which found a high level of agreement between the scores of e-rater and the human raters in high-stakes assessment contexts. However, some of the previous studies (Hoang & Kunnan, 2016; Huang, 2014) found similar results to those of the present study by revealing a difference between the scores assigned by e-rater and human raters. The discrepancy between e-rater scores and human rater scores might be due to the impact of essay length as previous studies indicated that e-rater can be influenced by word count (Attali & Burstein, 2006; Chodorow & Burstein, 2004). In the current study, the essays written on Topic Two were considerably longer than the essays on Topic One (average 421 and 342 words respectively). Therefore, e-rater might have given higher scores to the essays on Topic Two. In addition, although the raters received a detailed rater training on using the holistic scoring scale consistently, they might have focused on different aspects of writing from the aspects e-rater based its assessment on (Bauer & Zapata-Rivera, 2020; Li et al., 2014) or scoring differences might have occurred due to the limitations of computer-based scoring (Zehner et al., 2018). Moreover, the standard deviation of the holistic scores given by e-rater was lower than that of the holistic scores given by human raters for Topic One and Topic Two, indicating that the human raters assigned more various scores than e-rater to the same EFL essays.

The second research question investigated the sources of score variation contributing to the holistic scores obtained from human raters and the impact of e-rater on the sources of variation. G-theory analysis showed that the variance components (person, rater, and the residual) had similar portions of contribution to the total variance before and after e-rater scores were included in the analysis for both topics. For Topic One, rater was the third largest variance component (20.2 %) while for Topic Two it yielded the greatest variance (48.7 %). These results indicate that the human raters were inconsistent in terms of severity and leniency while scoring EFL essays even though they received a detailed rater training on how to apply the given scoring criteria in a consistent way before the scoring procedure. This contradicts the literature which suggested that rater consistency might be improved through rater training because training enables raters to have a clear conception regarding the quality of a piece of writing (Homburg, 1984; Shohamy et al., 1992). Two factors might have contributed to this result: First, the raters had not received any training on scoring essays before they participated in this study. One session of training might not have been sufficient to reconcile their rating behaviours. Second, because the raters were inexperienced in using the given scale, they might have displayed an inappropriate use of it or still used internal criteria (Barkaoui, 2010). The participants were accustomed to using a 100-point holistic scale which was designed by their institution considering the expectations of the institution from its students. Yet, they were required to use the 6-point holistic scale which was developed by ETS to score the writing tests in high-stakes standardized tests. Therefore, the raters might have needed a more intensive and ongoing training and calibration program on how to use the given scoring scale. The residual was the greatest source of score variation for Topic One (40.1 %) while it was the second largest variance for Topic Two (34.1 %). These results indicate that other facets which were not considered in the current design might have contributed to the score variance (Brennan, 2001; Huang et al., 2014). For example, the writing task (i.e., this study included only argumentative essays written on two different topics) and the quality of essays were not considered in the present study. However, previous studies showed that various writing tasks can affect the variability and reliability of scores in EFL writing assessment (Huang, 2008; Lee & Kantor, 2005), and that raters tend to give a lower score to an EFL essay when they think it has low level of quality with regards to simple construction and lexicon (Engber, 1995; Song & Carusa, 1996).

The third research question asked whether integrating e-rater scores with human rater scores impacted the scoring reliability in different rating scenarios. The results showed that higher generalizability and dependability coefficients were obtained when e-rater scores were integrated with the human rater scores. This result indicates that e-rater can replace one of the human raters in order to provide a more reliable and cost-effective writing assessment procedure. This result is in line with the current use of e-rater in the writing sections of GRE and TOEFL iBT tests, where e-rater could reliably replace one of the two human raters (Bridgeman et al., 2012). However, it should be noted that the human raters of the present study had not received any rater training before this study and they received only one session of training for this study. Therefore, it can be more realistic to conclude that e-rater can replace the raters who are not exposed to intensive rater training based on the results of the present study.

5. Conclusions and Implications

Overall, the results of the present study indicate that human raters show a considerable inconsistency in scoring the same EFL essays even if they are provided with a detailed rater training, which endangers the fairness of test scores. In order to attain more reliable scores and fairer judgments, literature suggests including two or more raters in the assessment process and assessing students' writing ability through various tasks or topics (Lee et al., 2002). The present study concluded that e-rater scores can be integrated with human rater scores in order to get more reliable ratings. As well as providing reliable ratings, integrating e-rater in the assessment procedure can be more practical and cost-effective (e.g., less time-taking and more labour-saving) than integrating two or more raters or using various tasks in writing assessment (Chodorow & Burstein, 2004; Shermis & Burstein, 2003; Shermis et al., 2010; Williamson et al., 2012). For practicality and cost-effectiveness, e-rater can be used reliably in high-stakes EFL writing assessment contexts (e.g., English proficiency exams for entrance and exit ELT departments or writing tests applied for selecting students for international exchange programs) in Turkish universities. Furthermore, it is suggested that EFL instructors are provided with a more intensive training program, which includes two or more sessions, before they score students' writing performance for high-stakes decisions. Training is important for EFL instructors because they do not generally receive training while scoring their students' writing tasks in authentic writing assessment contexts. If it is possible, the raters' use of the scale can be monitored for a short period (e.g., maybe in low-stakes assessment contexts) in order to ensure that they can apply the scale reliably (Harsch & Martin, 2012; Weigle, 2002); then, the raters who deviate from the norm can be retrained or excluded from the scoring procedure (Shohamy et al., 1992; Weigle, 2002). Additionally, selecting a clear and specific scoring scale can assist raters to assign more reliable scores (Shohamy et al., 1992). The raters should be involved in the scale selection procedure or they can develop their own scale in line with their specific purposes (Harsch & Martin, 2012).

This study has three limitations. First, this study used only argumentative essays written by EFL students to examine the variability and reliability of scores. Research has revealed that different writing tasks (e.g., argumentative essays and narrative essays) can affect the variability and reliability of EFL writing scores (Huang, 2008; Lee & Kantor, 2005). Second, the present study did not include qualitative data regarding the evaluation criteria and scoring processes of raters, which poses an obstacle to explain what criteria the raters based their judgements on while scoring the essays. Third, the raters of this study underwent only one session of training, which might not have been sufficient to calibrate the raters who had not received any training before. Therefore, future studies can use different writing tasks to investigate the variability and reliability of scores and collect qualitative data with the purpose of understanding how raters' assessment standards and expectations impact their holistic scoring behaviours. Moreover, future studies where the raters are provided with a more intensive training program can be conducted and the results can be compared to those revealed by the current study.

6. Authors' Note

This study was derived from a PhD dissertation by Elif SARI under the supervision of Turgay HAN and it was approved by the Faculty of Educational Sciences of Atatürk University in Turkey.

7. Acknowledgements

We would like to express our gratitude to TUBITAK (The Scientific and Technological Research Council of Turkey) for the financial support provided through 2211-National Graduate Scholarship Programme. We also owe our thanks to Educational Testing Services (ETS) for supporting this study by donating licenses to use Criterion Online Writing Evaluation Service.

8. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V. 2.0. *The Journal of Technology, Learning and Assessment, 4*(3), 3-30. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baker, B. A. B. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing, 15*, 133–153. http://dx.doi.org/10.1016/j.asw.2010.06.002

Barkaoui, K. (2010). Do ESL essays raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31-57.

Bauer, M. I., & Zapata-Rivera, D. (2020). Cognitive Foundations of Automated Scoring. In *Handbook of Automated Scoring* (pp. 13-28). Chapman and Hall/CRC.

Blood, I. (2011). Automated essay scoring: A literature review. *Studies in Applied Linguistics and TESOL*, *11*(2), 40-64.

Brennan, R. L. (2001). *Generalizability theory: Statistics for social science and public policy*. New York: Springer-Verlag. Retrieved from https://www.google.com.tr/search?hl=tr&tbo=p&tbm=bks&q=isbn:0387952829

Bridgeman, B., Trapani, C, & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27-40. https://doi.org/10.1080/08957347.2012.635502

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of Psychology, 52*(1), 13-15. http://dx.doi.org/10.1016/j.jsp.2013.11.008

Brown, H. D. (2004). *Language assessment: Principles and classroom practice*. New York, NY: Pearson/Longman.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., ... & Wolff, S. (1998). Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays. *ETS Research Report Series*, *1998*(1), i-67. http://dx.doi.org/10.1002/j.2333-8504.1998.tb01764.x

Chang, Y. (2002). EFL teachers' responses to L2 writing. Reports Research (143). Retrieved from http://files.eric.ed.gov/fulltext/ED465283.pdf on March 23, 2015

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (Research report No. 73). Princeton, NJ: Educational Testing Service.

Cronbach, L. J., Gleser G. C., and Nanda H. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley

Ebyary, K., &Windeatt, S. (2010). The impact of computer-based feedback on students' written work. *International Journal of English Studies, 10*(2), 121-142. https://doi.org/10.6018/ijes/2010/2/119231

Elliot, S. (2001). *Applying IntelliMetric Technology to the scoring of 3rd and 8th grade standardized writing assessments* (RB-524). Newtown, PA: Vantage Learning.

Elorbany, R., & Huang, J. (2012). Examining the impact of rater educational background on ESL writing assessment: A generalizability theory approach. *Language and Communication Quarterly, 1*(1)*,* 2-24.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of second language writing*, *4*(2), 139-155. https://doi.org/10.1016/1060-3743(95)90004-7

Foltz, P. W., Kintsch W., & Landauer, T. K. (1999). The measurement of textual coherence with Latent Semantic Analysis. *Organizational Process, 25*(2-3), 285-307. https://doi.org/10.1080/01638539809545029

Güler, N., Uyanık, G. K., & Teker, G. T. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi Yayınları.

Han, T. (2013). The impact of rating methods and rater training on the variability and reliability of EFL students' classroom-based writing assessments in Turkish universities: An investigation of problems and solutions. Atatürk University, Erzurum, Turkey.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, *17*(4), 228-250.

Heaton J. B. (2003). *Writing English language tests*. USA: Longman.

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly*, *13*(4), 359-376. https://doi.org/10.1080/15434303.2016.1230121

Hockly, N. (2019). "Automated Writing Evaluation". *ELT Journal, 73*(1), 82-88. https://doi.org/10.1093/elt/ccy044

Homburg, T.J. (1984). "Holistic Evaluation of ESL Composition: Can It be Validated Objectively?" *TESOL Quarterly,* 18(1), 87-108.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? - A generalizability theory approach. *Assessing Writing*, *13*(3), 201-218. http://dx.doi.org/10.1016/j.asw.2008.10.002

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

Huang, S. J. (2014). Automated versus Human Scoring: A Case Study in an EFL Context. *Electronic Journal of Foreign Language Teaching*, *11*.

Hyland, K. (2003). *Second language writing*. New York, NY: Cambridge University Press.

James, C. L. (2006). Validating a computerized scoring system for assessing writing and placing students in composition courses. *Assessing Writing*, *11*, 167-178. https://doi.org/10.1016/j.asw.2007.01.002

Johnson, R.L., Penny, J.A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* New York: The Guilford Press.

Kieffer, K. M. (1998, April). *Why generalizability theory is essential and classical test theory is often inadequate?* Paper presented at the Annual Meeting of the South Western Psychological Association, New Orleans, LA.

Latifi, F. S., & Gierl, M. J. (2020). Automated scoring of junior high essays using Coh-Metrix features: Implications for large-scale language testing. Language Testing. https://doi.org/10.1177/0265532220929918

Lee, Y. W., & Kantor, R. (2005). Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes. *ETS Research Report Series*, *2005*(1), i-76. https://doi.org/10.1002/j.2333-8504.2005.tb01991.x

Lee, Y.-W., Kantor, R., & Mollaun, P. (2002). "Score Dependability of the Writing and Speaking Sections of New TOEFL". [Proceeding]. *Paper Presented at the Annual Meeting of National Council on Measurement in Education*, New Orleans: LA. Abstract retrieved on December 11, 2012 from ERIC. (*ERIC No. ED464962)*

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, *44*, 66-78. https://doi.org/10.1016/j.system.2014.02.007

Lim, G. S. (2009). Prompt and rater effect in second language writing performance assesment. (Doctoral dissertation, The University of Michigan). Retrieved from http://deepblue.lib.umich.edu on March 23, 2015

Liu, S., & Kunnan, A. (2016). Investigating the Application of Automated Writing Evaluation to Chinese Undergraduate English Majors: A Case Study of WriteToLearn. *CALICO*, *33*(1), 71-91. https://doi.org/10.1558/cj.v33i1.26380

Popham, J.W. (1981). *Modern educational measurement.* Englewood: Prentice.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A premier.* Newbury Park, CA: Sage

Shermis, M. D., & Burstein, J. (2003). *Automated Essay Scoring: A cross disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 20-26). Oxford, UK: Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00233-5

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait rating for automated essay scoring. *Educational and Psychological Measures, 62*, 5-18. https://doi.org/10.1177/001316440206200101

Shi, L. (2001). Native- and Nonnative-Speaking EFL Teachers' Evaluation of Chinese Students' English Writing. *Language Testing, 18*(3), 303-325.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1)*,* 27-33. https://doi.org/10.1111/j.1540-4781.1992.tb02574.x

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking, and ESL students? *Journal of Second Language Writing, 5*(2), 163-182. https://doi.org/10.1016/S1060-3743(96)90023-5

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 1-24. https://doi.org/10.1191%2F1362168806lr190oa

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223. http://dx.doi.org/10.1177/026553229401100206

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

Williamson, D., Xi, X., & Breyer, J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement, Issues and Practice, 31*(1), 2-13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Zehner, F., Goldhammer, F., & Sälzer, C. (2018). Automatically analyzing text responses for exploring gender-specific cognitions in PISA reading. *Large-scale Assessments in Education*, *6*(1), 1-26.