International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 660–674
https://doi.org/10.21449/ijate.1384824
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Investigating the quality of a high-stakes EFL writing assessment procedure in the Turkish higher education context

Elif Sarı [1*]

[1]Karadeniz Technical University, School of Foreign Languages, Trabzon, Türkiye

ARTICLE HISTORY

Received: Nov. 01, 2023

Accepted: Aug. 26, 2024

Keywords:

EFL writing assessment,
Writing assessment in higher education,
Scoring variability,
Scoring reliability,
Generalizability (G-) theory.

Abstract: Employing G-theory and rater interviews, the study investigated how a high-stakes writing assessment procedure (i.e., a single-task, single-rater, and holistic scoring procedure) impacted the variability and reliability of its scores within the Turkish higher education context. Thirty-two essays written on two different writing tasks (i.e., narrative and opinion) by 16 EFL students studying at a Turkish state university were scored by 10 instructor raters both holistically and analytically. After the raters completed the scoring procedure, semi-structured individual interviews were held with them to gain insight into their views regarding the quality of the current scoring procedure. The G-theory results showed that the reliability coefficients obtained from the current scoring procedure would not be sufficient to draw sound conclusions. The quantitative results were partly supported by the qualitative data. Important implications were discussed to improve the quality of the current high-stakes EFL writing assessment policy.

1. INTRODUCTION

Reliability and validity are the two fundamental components of assessment. Reliability refers to the consistency of scores obtained across a range of circumstances and conditions (Johnson et al., 2009). Without consistency, it becomes challenging to draw meaningful conclusions or make accurate inferences about an individual's true ability. Validity, as the other important concept in assessment, refers to the degree to which an assessment tool accurately measures what it claims to measure (Bachman, 1990). It means that validity is "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (Messick, 1989, p. 39). If a given score is not valid, that would impair the fairness of the judgment made about the test takers' performance (Kane, 2010). Although consistency in test scores does not necessarily ensure validity, it is a fundamental requirement for it (Popham, 1981). Consequently, reliability is viewed "as a cornerstone of sound performance assessment" (Huang, 2008, p. 202).

It is necessary to ensure the reliability and fairness of scores in any assessment procedure, especially when the decisions made on these scores significantly impact students' lives (AERA, APA, & NCME, 2014). However, it is difficult to provide consistency among or within raters due to a variety of rater differences, such as educational background, linguistic background,

*CONTACT: Elif SARI ✉ elifsari@ktu.edu.tr Karadeniz Technical University, School of Foreign Languages, Trabzon, Türkiye

e-ISSN: 2148-7456

professional experience, and beliefs and expectations (Huot, 1990). The factors impacting the reliability and fairness of scores in ESL/EFL writing assessment can be categorized under three headings: 1) the factors related to the rater, 2) the factors related to the writing task, and 3) the factors related to the scoring method (Barkaoui, 2007; Barkaoui, 2008; Gebril, 2009; Huang, 2011; Weigle, 2002).

The literature has shown that rater-related factors such as the rater's native language (Cheong, 2012; Kim & Gennaro, 2012; Shi, 2001), professional experience (Barkaoui, 2010; Rinnert & Kobayshi, 2001; Şahan & Razı, 2020), professional background (Elorbany & Huang, 2012; Weigle, Boldt, & Valsecchi, 2003), and training (Attali, 2020; Fahim & Bijani, 2011; Weigle, 1994) affect the scoring variability and reliability. Several studies indicated that native English-speaking (NES) raters exhibited different scoring tendencies from non-native English-speaking (NNES) raters. Shi (2001) discovered that NES raters tended to exhibit a more favorable disposition when scoring content and language aspects, whereas NNES raters showed a tendency to be critical, particularly regarding organization and essay length. Similarly, in Kim and Gennaro's (2012) research, NNES raters were inclined to be more severe and displayed more variability in their scoring compared to NES raters. In contrast, Cheong (2012) observed that NES raters awarded lower grades and applied stricter evaluation criteria across three domains: content, organization, and language use. Regarding the impact of raters' professional experience on their scores, Rinnert and Kobayashi (2001) concluded that the least experienced Japanese raters gave higher scores compared to NES raters, and the groups differed in the criteria they prioritized. Barkaoui (2010) found that when employing holistic and analytic scales, experienced and inexperienced raters exhibited varying degrees of severity and leniency. Novice raters tended to be more lenient in their ratings compared to experienced raters. In addition, Şahan (2018) observed that highly experienced raters were more lenient and assigned higher scores, particularly for low-quality essays. To investigate how raters' professional backgrounds impact their scoring behaviours, Weigle, Boldt, and Valsecchi (2003) studied how instructors from different professional backgrounds evaluate text-responsible writing by ESL students. They found that raters from different disciplines had varying assessments, with English department raters being the strictest and history department raters being the most lenient. The study also revealed that English department raters placed more emphasis on grammar. In a separate study by Elorbany and Huang (2012), it was observed that raters with different educational backgrounds displayed different assessment behavious. Teacher candidates majoring in TESOL provided more consistent scores compared to the raters who did not have a TESOL background. To reveal the impact of rater training on raters' scores, Weigle (1994) studied experienced and inexperienced raters' scoring behaviours before and after they received training and revealed that inexperienced raters' scoring behaviours changed after training while the others gave similar scores both before and after the training. Similarly, Fahim and Bijani's (2011) study found that providing training to raters improved self-consistency and reduced severity and bias in the rating process. Finally, Attali (2020) compared inexperienced and experienced raters and found their ratings to be similar after initial training, but inexperienced raters showed more score variability.

Several studies indicated that writing task (e.g., narrative, persuasive, etc.) is another factor that affects the scoring variability and reliability (Cumming et al., 2002; Gebril, 2009; Hamp-Lyons & Mathias, 1994; Weigle, 1999; Zhao & Huang, 2020). For instance, as Hamp-Lyons and Mathias (1994) discovered, essays written in response to challenging writing prompts were given higher scores than those written in response to easy prompts. They also discovered that the category that the raters considered the simplest received the lowest ratings, whereas the category perceived as the most challenging received the highest ratings. In a similar vein, Weigle (1999) found that inexperienced raters assigned lower grades to certain essay types compared to experienced raters, but training reduced the differences. Cumming et al. (2002) also observed that writing tasks influenced raters' scoring processes and their focus on different

essay features. Additionally, Gebril (2009) and Zhao and Huang (2020) showed that including different task types increased scoring reliability.

The scoring method used by raters also affects the score variability and reliability in writing assessments. Therefore, several studies were undertaken to investigate how holistic and analytic scoring methods impact the variability and reliability of scores (Barkaoui, 2007, 2010; Han, 2013; Liu & Huang, 2020; Song & Caruso, 1996). For instance, in their study, Song and Caruso (1996) compared the holistic and analytic scoring of compositions written by native and non-native English speakers and found no statistically significant difference between the groups stemming from the rating method. Barkaoui (2007) investigated how different scoring methods impacted EFL essays and found higher inter-rater reliability with holistic rating. In a later study, Barkaoui (2010) examined the influence of the rating method on writing evaluation and found that the rating method significantly impacted the raters' scoring processes and the writing aspects they prioritized. In the same vein, Han's (2017) study suggested that detailed training made holistic scoring as reliable as analytic scoring. More recently, Liu and Huang (2020) evaluated the scoring policy of a standardized EFL assessment in China and showed that analytic scoring produced more reliable scores. It also showed that scoring reliability could improve with the increased number of tasks.

To sum up, the research has indicated that ESL/EFL writing assessment is a problematic issue as it is essential to control several factors that impact the variability, reliability, and thus the fairness of scores. In this sense, it is crucial to investigate the variability and reliability issues in any writing assessment procedure that is used to make critical judgments about the examinees' writing abilities (AERA, APA, & NCME, 2014). For example, in Turkish higher education, students' writing performance is assessed to make some high-stakes decisions such as determining students' language proficiency when they are enrolled in the departments that are related to English Language Teaching or Literature or selecting students who will take part in the international exchange programs like Erasmus+. Although each university conducts its own writing assessment procedure, students' writing performance is mostly assessed using a single-task, single-rater, and holistic scoring procedure as it is considered to be more time-efficient and cost-effective. Since the studies reviewed above were mostly conducted in different writing assessment contexts, there is limited information regarding the scoring variability and reliability of the writing assessment procedures employed specifically in the context of Turkish higher education. Therefore, it becomes imperative to undertake an in-depth exploration of the quality of writing assessment within this specific educational context. To bridge this existing gap in the literature, this study set out to evaluate the quality of a single-task, single-rater, and holistic scoring method within the Turkish higher education context, focusing on its potential effects on scoring variability and reliability using the G-theory framework. Studying the variability and reliability of this institutional writing evaluation process can have significant implications for assessment policymakers in this specific context (i.e., the Turkish higher education context) as it helps them determine the optimal approach for a high-quality writing assessment procedure, focusing on key factors such as the number of tasks, the number of raters, and the scoring method. Furthermore, the implications are far-reaching and extend to professionals engaged in the evaluation of EFL writing skills on a global scale. Consequently, the findings and insights generated by this study could substantially inform and enhance the practices and policies of assessment professionals and policymakers alike, with the potential to foster improvements not only in Turkish higher education but also in the broader context of EFL writing assessment. The study was directed by four specific research questions, which are as follows:

1. What are the sources of variability in scores given to the EFL papers?
2. How reliable are the EFL scores in terms of G-coefficients for norm-referenced interpretation and dependability coefficients for criterion-referenced score interpretations?

3. Does the scoring reliability change when the number of raters, tasks and the scoring method change?

4. What are the raters' views regarding the overall quality of the single-task, single-rater, and holistic writing assessment procedure?

1.1. G-theory Framework

Classical Test Theory (CTT), the conventional measurement model, posits that a measured score (X) comprises a true score (T) and an error score (E). The true score is the test-takers' actual performance resulting from their ability, while the observed score reflects the interaction between the true score and the error score, which are influenced by some external factors apart from the ability intended to be measured (Fulcher & Davidson, 2007). CTT primarily considers two sources of error (i.e., a single ability and a single source of errors), while G-theory recognizes that the sources of error in measurement are diverse and can come from various facets or components. (Bachman, 1990; Briesch et al., 2014). These sources, commonly known as facets, can include different raters, items, occasions, or any other factors that contribute to measurement variability. By incorporating these facets into the analysis, G-theory provides a more detailed understanding of how these different sources impact the reliability and generalizability of the obtained scores. (Shavelson & Webb, 1991).

The G-theory analysis includes two phases: the generalizability study (G-study) and the decision study (D-study). The G-study focuses on assessing the generalizability, or the extent to which the obtained results can be applied beyond the specific conditions of the study. It aims to estimate the various sources of error in measurement and to determine how they contribute to the variability of scores. By examining different facets of measurement, such as raters, tasks, and occasions, the G-study helps researchers understand the factors that affect the reliability and validity of the measurement instrument or procedure (Barkaoui, 2007; Huang et al., 2014). The D-study, on the other hand, is a phase that focuses on making decisions using the measurement data revealed in the G-study. By utilizing the results from the G-study, which provides insights into the various sources of error and their contributions to score variability, the D-study aims to optimize measurement practices and evaluate the reliability of the proposed procedures. (Keiffer, 1998; Huang, 2008). The D-study is essential for determining the adequacy of the measurement procedure for the specific decision-making context as it allows researchers to determine which facets or factors should be prioritized for improvement or control in assessment procedures. (Briesch et al., 2014). Overall, the D-study extends the findings of the G-study by guiding how to improve measurement procedures.

G-theory was employed as the theoretical framework of the quantitative analyses in this study because of its sophisticated and robust nature in the field of ESL/EFL writing assessment. The primary goal was to explore the intricate interplay of several key factors within the assessment process: the number of raters, the variety of tasks presented to the students, and the specific assessment methods employed. In doing so, the study aimed to shed light on how these multifaceted elements collectively influence the variability and reliability of an institutional high-stakes EFL writing assessment procedure.

2. METHOD

The present study is a descriptive research as it aims at describing the existing situation without manipulating the variables and making the necessary determinations based on the data obtained. This descriptive study incorporated both quantitative and qualitative data to answer the research questions. The quantitative data were collected to find out the variability and reliability of scores obtained from this specific assessment procedure while the qualitative data were collected to search out the raters' perspectives of the scoring procedure.

2.1. Selection of Writing Samples

The writing samples of this study were collected from the School of Foreign Languages at a Turkish state university in the 2022-2023 academic year. Forty-five B1-level students (19 female and 26 male, aged 18 to 24) from the English preparatory program were required to write two essays in separate sessions, as it is impossible to assess task effects using a single-task scenario within the G-theory framework. In the first session, the students were required to write a narrative essay on *"Write about your worst, best, or most embarrassing time in your life"*. In the second session, they were tasked to write an opinion essay on *"Write about advantages and disadvantages of living in a big city"*. The topics were selected from the institutional English proficiency exams administered in the previous years. Following the same procedure administered in the institutional exams, the students were required to write each of their 200-220 word essays in 30 minutes using pen and paper. Totally 90 essays were collected from the students. Then, to ensure a wide range of variation among the essays, two independent raters, who did not participate as raters in the scoring procedure of the study, meticulously categorized the essays into three qualities (i.e., high, medium, and low) using the holistic scoring scale used in the scoring procedure. The raters did not assign numerical scores to the essays during this process. Only the essays that were consistently classified as having either high- or low-quality by both raters were selected for further analysis. As a result, a total of 32 essays, written by 16 students, were determined to be used as the sample for the current study.

2.2. Selection of Raters

The purposive convenience sampling method was used to select the EFL instructors based on their willingness to volunteer their time and their proximity to the researcher (Creswell, 2012). The raters had to meet the following criteria: a) being a full-time employee at an EFL teaching institution, b) having experience in teaching EFL writing, and c) having participated in the institutional high-stakes writing assessment. As a result, ten instructors, consisting of six females and four males, took part in this study as raters. They were highly skilled in EFL teaching, boasting expertise in teaching and assessing writing with at least ten years of experience. The instructors were full-time employees of a Turkish state university and native Turkish speakers, with ages ranging between 36 and 52 with a mean of 43. All of the raters were informed about the purpose of the study and they wholeheartedly agreed to participate in the study. To ensure privacy and confidentiality, the participants' identities were kept confidential through the use of pseudonyms.

2.3. Scoring Rubrics

One of the primary objectives of this study is to investigate how the choice of scoring method impacts the variability and reliability of scores. To achieve this, the raters were tasked with evaluating the essays twice, employing two different approaches: initially utilizing a holistic method, followed by an analytical approach, with a three-week time interval. The holistic scale was the authentic institutional scale used for the high-stakes writing assessment, which required the raters to assign a single overall score, out of 100 points, to an essay based on its content and organization, language use, and mechanics. An adapted version of the analytic scale Jacobs et al. (1981) developed was used in analytic scoring because its scoring criteria were compatible with those of the holistic scale, but this time they were required to assign a score for each of the five categories: a) content (30 pts.), b) organization (20 pts.), c) grammar (20 pts.), d) vocabulary (20 pts.), and e) mechanics (10 pts.).

2.4. Scoring Procedure

Before the scoring procedure, the raters were thoroughly informed of the purpose of the study and presented with a consent form ensuring the protection of their rights and the confidentiality of the obtained data. Following this, the raters were introduced to the holistic scale, and they assessed three essays representing different proficiency levels (low, medium, and high) to build

a common understanding of the scoring criteria they used. They discussed the differences in their scores to align their expectations and judgments. Then, the raters were given a set of materials, which included 32 essays on two different topics, one holistic scoring rubric, one scoring form to write the scores on, and a questionnaire that was formed to gather background information about the raters. Three weeks after they completed holistic scoring, they were introduced to the analytic scale. The three-week time interval was set to prevent paper familiarity. The components of each level on the scale and what they signified were explained until the expectations were all clear. Once again, the raters evaluated three essays representing varying proficiency levels analytically and discussed the disparities in their scores. Finally, the raters were required to score the 32 essays analytically. The raters did not receive extensive training for holistic and analytic scoring in this study, as they had already been trained in assessing institutional exam papers.

2.5. Interviews with Raters

After completing both the holistic and analytic scoring procedures, all raters were interviewed individually to gather their perceptions of the single-task, single-rater, and holistic scoring methods used in their institution. Each interview lasted nearly 15 minutes with four main questions regarding the number of writing tasks, the number of raters, the scoring method, and the current assessment procedure in general. Some extra questions were asked when it was felt necessary to get further explanation on the answers. The interviews were carried out in Turkish to gather more detailed information. The interviews were recorded, transcribed, and then translated into English by the author of this study, which were checked by another researcher who had experience in analysing qualitative data.

2.6. Data Analysis

This study utilized the G-theory framework to analyze quantitative data to investigate the influence of various factors such as paper, task, rater, and their interactions on the variance of scores obtained from holistic and analytic scoring using the EduG computer program. The researcher conducted two distinct G-studies, one dedicated to holistic scoring and the other to analytic scoring. Each of these G-studies took into account the random effects of the combination of individuals, tasks, and raters, denoted as person-by-task-by-rater (p×t×r). By separately analyzing holistic and analytic scoring, the study aimed to gain a nuanced understanding of how these different approaches contribute to score variance, shedding light on their specific strengths and weaknesses. Furthermore, the research delved into a separate realm of analysis through two random effects D-studies, one for each scoring method (holistic and analytic). These D-studies were conducted to calculate generalizability coefficients, which are typically used in norm-referenced tests to assess the extent to which assessment outcomes can be generalized, and dependability coefficients, which are employed in criterion-referenced tests to gauge the reliability of the assessment process. The D-studies were executed with varying numbers of raters and tasks, offering insights into the impact of these key variables on the reliability and validity of the scoring methods. The culmination of these analyses not only enriched our understanding of the assessment processes but also furnished valuable insights for future test design and evaluation practices.

Furthermore, the qualitative data obtained through the rater interviews were analysed through manual content analysis as suggested by Creswell (2012). The author of this study compiled the student answers under each interview question. The author proceeded to conduct a more in-depth examination of the compiled student answers. The data were carefully scrutinized, and similar responses were grouped together under specific categories. This process was carried out by both the author and another experienced researcher, who worked independently to ensure that their categorization was unbiased. Then, the author and the researcher worked together to sort the categories into themes that corresponded with the interview questions. Direct quotes from the interviews were also included to increase the validity of the qualitative data.

2.7. Validity and Reliability of Data Collection Tools and Procedure

To ensure the reliability and validity of both the data collection tools and procedures, several precautions were implemented. First, students generated writing samples under conditions mirroring those of the actual institutional writing exams, with topic selection based on real exam topics tailored to students' proficiency levels and familiarity. Second, two independent raters categorized the collected writing samples into high, medium, and low qualities and the papers which the two raters agreed to be high-quality or low-quality were selected for data analysis. Third, the raters were introduced to the criteria of holistic and analytic rubrics before the scoring procedure. They individually scored three sample essays using these rubrics and engaged in discussions until a consensus was reached on their understanding of the criteria and expectations. This aimed to minimize inconsistencies arising from potential misunderstandings. In addition, a three-week interval was introduced between the holistic and analytic scoring procedures to mitigate rater familiarity with the papers. Finally, to enhance the reliability of qualitative data analysis, the author collaborated with another experienced researcher during the qualitative data analysis procedure.

3. RESULTS

3.1. The Results of Random Effects Person-by-task-by-rater (p × t × r) G-studies

Specifically, two distinct random effects G-studies, one focusing on holistic scores and the other on analytic scores, were conducted. These G-studies allowed us to scrutinize the multifaceted factors contributing to the overall variance observed in the scoring of the 32 papers. The assessment encompassed a person-by-task-by-rater (p × t × r) framework, which means that we explored how individual students, the specific tasks assigned, and the raters who assessed the papers collectively influenced the final scores. By doing so, we were able to unravel the complex web of interactions among these key components, shedding light on the various aspects that impacted the overall variance in the scoring process. The outcomes of these analyses are given in Table 1.

Table 1. *Variance components for random effects p × t × r G-study.*

Type of Scores	Source of Variability	*df*	σ^2	%
Holistic Scores	*p*	15	.55	20.8
	t	1	.10	3.8
	r	9	.50	19.9
	pt	15	.82	30.8
	pr	135	.10	4.1
	tr	9	.16	6.1
	ptr	135	.65	24.6
	Total	319	2.63	100
Analytic Scores	*p*	15	.99	38.9
	t	1	.02	0
	r	9	.26	9.8
	pt	15	.23	9.1
	pr	135	.09	3.9
	tr	9	.04	1.7
	ptr	135	.67	26.5
	Total	319	2.53	100

The breakdown of variance components for the holistic scoring, as presented in the Table 1, revealed that the largest contributor to the overall variance was the person-by-task (pt) interaction, accounting for a substantial 30.8% of the total variance. This outcome implies that the 16 EFL students exhibited significantly divergent performance levels in their execution of

the first and second writing tasks. The disparities in their output underscore the distinct challenges posed by these tasks, rendering them non-uniform in nature. Following closely, the residual component (ptr) emerged as the second most influential source of variance, representing 24.6% of the total variance. This component suggests that factors beyond the anticipated interactions among raters, writing tasks, and individual students played a significant role in the variations observed in the scores. These unexplained sources may encompass systematic and random errors, as well as latent factors that eluded detection in the present analysis, thereby underlining the multifaceted and nuanced nature of the holistic scoring process. Person (p) contributed 20.8% of the overall variance, signaling that the evaluation scores assigned to the 16 students were substantially shaped by their characteristics and competencies. These unique traits and skills held a discernible sway over the final scores, reinforcing the idea that the students' inherent abilities were integral to the assessment process. Additionally, the rater component, which represented 19.9% of the total variance, exhibited the raters' varying degrees of leniency or severity in their holistic marking of the papers. In essence, this suggests that the diversity in final scores could be attributed, to a considerable extent, to the idiosyncratic scoring tendencies of the raters. The task-by-rater (tr) component, at 6.1% of the total variance, hinted at the presence of considerable inconsistency among the raters in their evaluation of the two writing tasks. This inconsistency indicates that the raters had differing interpretations of the scoring criteria, further underscoring the intricate nature of the evaluation process. Meanwhile, the person-by-rater (pr) component contributed 4.1% of the total variance, emphasizing that the raters displayed inconsistencies in their evaluation of the essays authored by the 16 EFL learners who participated in this study. This irregularity points to a degree of subjectivity and variation in the raters' judgments. Finally, the task (t) component, representing 3.8% of the total variance, revealed a minor disparity in terms of the difficulty levels of the two tasks. This finding highlights that the tasks were not entirely equivalent in their demands, adding complexity to the holistic scoring process.

The breakdown of analytic scoring components, as outlined in Table 1, showed that the person (p) factor emerged as the most prominent contributor to the total variance, comprising a substantial 38.9%. This observation underscores a crucial point that the analytic scoring approach effectively discriminated among the 16 EFL learners, revealing significant disparities in their respective writing skills. Concurrently, the residual component (ptr), representing unexplained sources of variance, constituted the second-largest share of the total variance at 26.5%. This component serves as a critical reminder that not all aspects of scoring variability can be accounted for, highlighting the inherent complexity of the assessment process. Another salient finding was the rater (r) factor, which accounted for 9.8% of the total variance. This suggests that the raters themselves exhibited discernible differences in their approach, with some demonstrating greater leniency while others leaned towards severity when evaluating the papers analytically. This variance in rater behavior re-emphasizes the importance of consistency among raters in the assessment process. Moreover, the interaction between person and task (pt) contributed to 9.1% of the total variance, indicating that the nature of the writing tasks had a discernible influence on how raters approached analytic scoring. This finding highlights the need to consider the specific writing tasks and their inherent challenges when interpreting the assessment results. The person-by-rater interaction (pr) and task-by-rater interaction (tr) made up 3.9% and 1.7% of the total variance, respectively. These components highlight the complexity of the assessment process, where the interactions between individual learners and raters, as well as between writing tasks and raters, introduce additional layers of variability that can affect the final scores. Interestingly, the task (t) component accounted for 0% of the total variance, indicating that the difficulty of the writing tasks did not influence the raters' analytic scoring. This finding suggests a degree of consistency in the raters' approach across different writing tasks, despite the disparities in individual task complexities.

3.2. The Results of Person-by-task-by-rater (p × T × R) Random Effects D-studies

In order to thoroughly examine the reliability of the scores, we conducted two separate D-studies for holistic and analytic scoring, respectively. These D-studies were performed in a person-by-task-by-rater (p × T × R) framework, which means that we took into account variations across different individuals, tasks, and raters. The generalizability coefficient (Ep2) provides insights into the overall consistency and generalizability of the scores, helping us understand how reliably they can be applied in a broader context. The dependability coefficient (ϕ) allowed us to gauge the stability and dependability of the scores within the specific context of our analysis. By conducting these two distinct D-studies for both holistic scoring and analytic scoring, we aimed to understand the reliability and consistency of the scoring methods, which is vital for ensuring the accuracy and validity of our assessment process. The coefficients that are equal to or above 0.70 provide evidence that the scores are consistent and reliable measurements of the writing quality being assessed. The results of the D-studies are presented in Table 2.

Table 2. *Generalizability and dependability coefficients.*

Number of Papers	Number of Tasks	Number of Raters	Holistic Scoring		Analytic Scoring	
			Ep2	ϕ	Ep2	ϕ
16	1	1	.26	.21	.50	.39
16	1	2	.32	.27	.62	.53
16	1	3	.34	.30	.67	.60
16	1	4	.35	.31	.70	.64
16	1	10	.38	.35	.76	.73
16	2	1	.40	.31	.64	.48
16	2	2	.47	.39	.75	.62
16	2	3	.50	.43	.79	.69
16	2	4	.52	.46	.82	.74
16	**2**	**10**	**.55**	**.51**	**.86**	**.82**
16	3	1	.48	.37	.71	.52
16	3	2	.56	.47	.81	.66
16	3	3	.59	.52	.84	.73
16	3	4	.61	.54	.86	.77
16	3	10	.64	.60	.90	.86

For holistic scoring, as presented in Table 2, the generalizability and dependability coefficients in the current scenario involving 16 essays, two tasks, and ten raters were .55 and .51, respectively. In the single-task, single-rater, and holistic scoring procedure, the generalizability and the dependability coefficients would be .26 and .21, respectively, which would fail to reach the acceptable reliability coefficient of .70. This suggests that relying on a single rater and single task for scoring would result in lower reliability, indicating reduced generalizability of the scores to a larger population. If the number of raters and writing tasks was increased to two in this scenario, the generalizability and the dependability coefficients would be .47 and .39, respectively, which are far below the acceptable reliability coefficient of .70.

For analytic scoring, also given in Table 2, the generalizability and dependability coefficients in the current scenario involving 16 essays, two tasks, and ten raters were .86 and .82, respectively, which are significantly higher than the coefficients obtained from the holistic scoring. If analytic scoring was used in the single-task and single-rater scenario, the generalizability and the dependability coefficients would be .50 and .39, respectively, which are still below the acceptable reliability coefficient of .70 although they are much better than the coefficients obtained from the holistic scoring in the same scenario. If the number of raters and writing tasks was increased to two and analytic scoring was used instead of holistic scoring,

the generalizability and the dependability coefficients would increase to .75 and .62, respectively.

3.3. The Findings of the Rater Interviews

To gather the raters' views regarding the overall quality of the current institutional writing assessment procedure, four main questions were asked to the raters in the interviews held after they completed the scoring procedure. The analysis of the data obtained from the rater interviews yielded the following three themes that are related to each interview question: a) using a single writing task is sufficient in assessing students' writing skills; b) using a single rater is not appropriate for high-quality writing assessment; c) analytic scoring method provides more reliable results than holistic scoring method.

First, most of the raters stated that using a single writing task was sufficient in assessing EFL learners' writing skills. Contrary to what is suggested in the literature and what was found as a result of the random effects of person-by-task-by-rater D-studies conducted in the current study, the raters believed that increasing the number of writing tasks would not affect the score reliability. They commented that if the examinees were required to write two tasks, they would get more stressed and tired, which in turn would impact their performance negatively. In addition, they commented that scoring two tasks would not be practical in the high-stakes writing assessment context since a large number of examinees take this test and the results have to be announced in an expeditious manner. Only two of the raters suggested that if the number of writing tasks was increased from one to two, more reliable scores could be achieved.

Second, all of the raters agreed that using a single rater was not appropriate to provide a high-quality writing assessment procedure. They suggested that it is necessary to involve at least two raters in the scoring procedure for reliable and fair results in any high-stakes writing assessment contexts. Regarding this issue, one of the raters reported that *"As raters differ from each other in terms of their scoring behaviours, some raters tend to give high scores while the others tend to give low scores. Therefore, involving two raters in the scoring procedure was effective in decreasing the measurement error stemming from raters' tendencies"*. They also suggested that when the gap between the two raters' scores is large, a third rater should be asked to score the same essay to increase the reliability. In addition, they argued that their scoring performance should be monitored periodically and they should be provided with some feedback regarding their performance. Moreover, they added that the institution should organize more detailed rater training programmes to improve the consistency among the instructor raters.

Finally, it became evident that a significant majority, specifically eight out of the ten raters, agreed that the holistic scoring approach was unsuitable due to concerns regarding score consistency and reliability. They believed that analytic scoring would yield more realistic scores as the rater had to read the essay again and again in order to decide its quality based on the detailed criteria given in the analytic scale. Based on their experiences of scoring the essays for this study, two of the raters made the following comments regarding this issue: *"I could decide the holistic scores after reading the essay only once, but while I was scoring the same essays analytically, I had to read them again to decide the score for each subcategory of the analytic scale (i.e., content, organization, grammar, vocabulary, and mechanics)"*, *"I had to think more about the details regarding organization, grammar, vocabulary, and mechanics while scoring the essays analytically, which made me think that my analytic scores were more accurate than the holistic scores I assigned to the same papers"*. In addition, one of the raters reported that *"I realized that I do not consider mechanics when I score an essay holistically"*. Another rater made the following comment: *"I realized that in holistic scoring the use of language is the component that impacts my score most. If the student can use the grammatical structures accurately, I tend to give a high score even if the content is not sufficient"*. However, another rater stated that *"Content is the most important quality for me while scoring an essay holistically. If the student can explain the topic adequately with necessary supporting details, I*

do not care about grammatical problems. However, the analytic scale prevented me from ignoring the other components that are necessary for high-quality writing". These comments show that the raters demonstrate varying scoring behaviours in holistic scoring, which might increase the variability of scores and thus decrease the score reliability. However, the analytic scale enabled them to consider each subcategory thoroughly while scoring the essays. In addition, the analytic scale limited their overgeneralization of a single aspect of writing. However, a contrasting perspective was voiced by two out of the ten raters who argued that holistic scoring might be a more suitable approach for the high-stakes writing assessment conducted within the institution centering on the belief that holistic scoring proved to be a more time-efficient method as compared to the analytic scoring system.

4. DISCUSSION and CONCLUSION

The study utilized G-theory and conducted interviews with raters to explore the influence of a single-task, single-rater, and holistic scoring approach on score variability and reliability within the Turkish higher education context. It was expected that the findings, while specific to this study, could offer valuable insights to assessment experts in various educational institutions. These insights would serve as a blueprint for them to reevaluate and enhance their own writing assessment procedures, particularly in terms of improving score consistency and reliability, extending the potential impact of this research beyond its immediate context.

First, the random effects of person-by-task-by-rater G-studies provided insights into the distribution of variance for the two scoring methods. The results showed that in analytic scoring the variance component attributed to individual persons, defined as the desired variance by Brennan (2001), constituted a substantially larger portion than in holistic scoring. This suggests that analytic scoring was more effective in distinguishing the EFL learners in terms of their writing skills compared to holistic scoring. In the present study, the undesired variance stemming from factors such as the rater, the interaction between individuals and raters, and the tasks and raters (Brennan, 2001) was larger in holistic scoring than it was in analytic scoring. Specifically, the variance attributed to the interaction between the task and raters was over three times greater for holistic scoring than it was for analytic scoring. In line with previous research (e.g., Cumming et al., 2002; Gebril, 2009; Zhao & Huang, 2020), this result indicated that the nature of the task influenced the raters' scores. In the present study, holistic scoring exhibited a greater task effect compared to analytic scoring. Additionally, in holistic scoring the variance associated with the rater accounted for nearly twice as much of the total variance compared to analytic scoring, suggesting that raters exhibited greater inconsistency in their evaluations when employing holistic scoring, particularly in terms of leniency or severity in their ratings. This finding is consistent with prior studies conducted by Barkaoui (2008) and Liu and Huang (2020), but it contradicts the results of Barkaoui's (2010) study, which indicated more rater inconsistency in holistic scoring. Moreover, the variance component referred to as residual, which encompasses the interaction between raters, writing tasks, individuals, and other unexplained systematic and unsystematic sources of error, significantly contributed to score variance in both scoring methods. This underscores the importance of carefully considering and standardizing scoring procedures to minimize measurement errors, as emphasized by Brennan (2001) and Huang et al. (2012).

Second, the person-by-task-by-rater random effects D-studies revealed that the score reliability coefficients obtained from the single-task, single-rater, and holistic scoring procedure would fall significantly short of meeting the acceptable reliability standards for holistic scoring. In contrast, analytic scoring showed more acceptable reliability coefficients. If two writing tasks and two raters were involved in the same assessment procedure, the reliability coefficients would still be lower in the holistic scoring, but in the analytic scoring, the reliability would reach an acceptable level in the norm-referenced assessment while it would be lower in the criterion-referenced assessment. These results revealed that, in accordance with existing

research (Lee et al., 2002; Liu & Huang, 2020; Zhao & Huang, 2020), increasing the number of raters and writing tasks would have a positive impact on the reliability coefficients in both holistic and analytic scoring methods. However, it's important to note that even with these improvements, holistic scoring would not reach satisfactory reliability coefficients. On the other hand, by opting for analytic scoring and concurrently increasing the number of raters and tasks, the assessment process would have a significant enhancement in terms of score reliability. In summary, the results suggest that while holistic scoring benefits from more raters and tasks, switching to analytic scoring would result in notably improved score reliability.

Finally, the findings obtained from the rater interviews showed that the raters were mostly positive about using a single-task in the high-stakes writing assessment procedure because they thought it was more practical and time-efficient in such an assessment context where a large number of examinees' papers must be scored in a short time. In addition, contrary to what the literature suggested and the quantitative results of this study showed, they believed that a single writing task would be sufficient to measure the EFL learners' writing performance. On the other hand, in line with what the literature suggested (Gebril, 2009; Weigle, 2002), the raters did not favour using a single rater in the assessment of high-stakes writing tests as it would endanger the reliability and fairness of the scores. They believed that involving two raters in the scoring procedure can provide more reliable scores. Further, the raters were mostly positive about the analytic scoring method giving the reason that it would yield more realistic and reliable scores because when scoring the essays analytically, they were to abide by the criteria specified in the scale rather than making decisions based on their personal judgments, as supported by the related literature (Barkaoui, 2008; Barkaoui, 2010). Further, in line with the literature (Attali, 2020; Fahim & Bijani, 2011; Weigle, 1994), they commented that receiving rater training periodically might alleviate the inconsistencies stemming from different rater behaviours.

Overall, the results of this study demonstrated that the single-task, single-rater, and holistic scoring procedure would not be sufficient to guarantee high-quality in terms of reliability and fairness issues. Since writing scores are used for making important decisions about examinees in Turkish higher education, it is crucial to make some revisions in the single-task, single-rater, and holistic scoring procedure in order to ensure low variability and high reliability of scores. For this reason, in light of the findings of this study, it is suggested that examinees are required to write at least two writing tasks, and these tasks are scored by at least two raters employing the analytic scoring method. Including a third rater in the scoring procedure when the gap between the two raters is large, might also be a solution to increase the score reliability. In addition, instructor raters must be provided with training for the implementation of the revised assessment procedure. They should be monitored at regular intervals and given feedback about their scoring performance. The assessment policy makers in the Turkish higher education context should consider these suggestions while designing the EFL writing assessment procedures to attain sound and reliable results and make appropriate improvements in EFL education provided in Turkish higher education. Following these suggestions can guarantee the quality of high-stakes writing assessment procedures.

It's essential to recognize two limitations of this study when interpreting its results. Firstly, the study was not carried out in a real high-stakes writing assessment environment, meaning that the data collected may not precisely mirror what occurs in an authentic setting. Raters and examinees might respond differently under the pressures and conditions of a genuine test. Secondly, the relatively small number of selected papers used in this study could restrict the generalizability of the findings to a broader context. To enhance the generalizability of these findings, future research should encompass a broader selection of papers and diverse EFL writing assessment scenarios within Turkish higher education. This will enable a more comprehensive understanding of the factors in different contexts.

Acknowledgments

I would like to express my gratitude to my colleagues, who wholeheartedly participated in this study as raters. Thank you all for your invaluable time.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author. **Ethics Committee Number:** Karadeniz Technical University, Ethics Committee for Social and Human Sciences, E-82554930-050.01.04-421870.

Orcid

Elif Sarı https://orcid.org/0000-0002-3597-7212

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. (2020). Effect of Immediate Elaborated Feedback on Rater Accuracy. *ETS Research Report Series*, *2020*(1), 1-15.

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

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107. https://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* [Unpublished doctoral dissertation, University of Toronto, Canada].

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

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

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

Cheong, S.H. (2012). Native-and nonnative-English-speaking raters' assessment behavior in the evaluation of NEAT essay writing samples. *영어교육연구*, *24*(2), 49-73.

Creswell, John W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4 th ed.). Pearson Education.

Cronbach, L.J., Gleser, G.C., Nada, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96.

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

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *International Journal of Language Testing*, *1*(1), 1-16.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all?. *Language Testing*, *26*(4), 507-531.

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

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing* (pp. 69-87). United Kingdom: Cambridge University Press. https://doi.org/10.1017/CBO9781139524551.009

Hamp-Lyons, L., & Mathias, S.P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*(1), 49-68. https://doi.org/10.1016/1060-3743(94)90005-1

Han, T., & Huang, J. (2017). examining the ımpact of scoring methods on the ınstitutional EFL writing assessment: A Turkish perspective. *PASAA: Journal of Language Teaching and Learning in Thailand*, *53*, 112-147.

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. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal, 2*(4), 423-443. https://doi.org/10.5054/tj.2011.269751

Huang, J., Han, T., Tavano, H., & Hairston, L. (2014). Using generalizability theory toexamine the impact of essay quality on rating variability and reliability of ESOL writing. In J. Huang & T. Han (Eds.), *Empirical quantitative research in social sciences: Examining significant differences and relationships*, (pp. 127-149). Untested Ideas Research Center.

Huot, B.A. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41,* 201-213. https://www.jstor.org/stable/358160

Huot, B. (2002). *(Re)Articulating writing assessment: Writing assessment for teaching and learning.* Logan, Utah: Utah State University Press.

Jacobs, H.J., Zingraf, S.A., Wormuth, D.R., Ha rtfiel, V.F., &Hughey, J.B.(1981). Testing ESL composition: A practical approach. Massachusetts: Newbury House.

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

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177-182.

Kieffer, K.M. (1998). *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.

Kenyon, D. (1992, February). *Introductory remarks at symposium on development and use of rating scales in language testing*. Paper presented at the 14th Language Testing Research Colloquium, Vancouver, British Columbia.

Kim, A.Y., & Gennaro, D.K. (2012). Scoring behavior of native vs. non-native speaker raters of writing exams. *Language Research*, *48*(2), 319-342.

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)*

Liu, Y., & Huang, J. (2020). The quality assurance of a national English writing assessment: Policy implications for quality improvement. *Studies in Educational Evaluation*, *67*, 100941.

McNamara, T.F. (1996). *Measuring second language performance.* Addison Wesley Longman.

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

Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *The Modern Language Journal, 85*, 189-209.

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

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325. https://doi.org/10.1177/026553220101800303

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*, 163-182.

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors?. *Language Testing*, *37*(3), 311-332.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. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. Assessing Writing, *6*(2), 145-178.

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

Weigle, S.C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL writing: A pilot study. *TESOL Quarterly, 37*(2), 345-354.

Zhao, C., & Huang, J. (2020). The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: Implications for assessment policy makers. *Studies in Educational Evaluation*, *67*, 100911.