2023, 2(2), 7-17

A Comparison of Analytic and Holistic Scales in the Evaluation of Learner Essays

Buket Demirbüken [a1], Büşra Ulu [a2]

[a1] *Marmara University, İstanbul, Türkiye*
[a2] *Gebze Technical University, İstanbul, Türkiye*

Abstract
This study aimed to investigate the strengths and weaknesses of two holistic and two analytic scales designed for evaluating written work. For this purpose, the holistic scales by American Council on the Teaching of Foreign Languages (ACTFL) (2012) and TOEFL (developed by Educational Testing Services, as cited in Hughes, 2009) were chosen while the analytic scales used in this study were developed by Anderson (as cited in Hughes, 2009) and Jacobs et al.'s (1981, as cited in Hughes, 2009) scoring profile. The data used in this study was collected from 15 preparatory school students who are enrolled in various departments in a public university in Türkiye. The data was analyzed by three raters who are instructors at different universities, which are also located in Türkiye. Each rater analyzed every paper four times by using the scales. The analysis of the data revealed the strengths and weaknesses of the selected holistic and analytic scales used in this study. The pros and cons of each scale in terms of evaluating the written works of students were then compared for both holistic and analytic scales. The interrater reliability was also calculated for each scale. The results of the analysis revealed that holistic scales were more practical than analytic ones in terms of applying on-site while analytic scales give more detailed and safer results and enable the possibility to give comprehensive feedback. The study concluded that the purpose of the exam and the rater is an important determinant for the choice of scale.

Keywords: Testing writing, analytic scales, holistic scales, interrater reliability.

Received Date: September 1st, 2023. Acceptance Date: December 2nd, 2023.

1. Introduction

Testing is an important part of the teaching and learning process and making this process objective enough is the responsibility of the test designers and teachers who conduct the tests for more reliable and accurate results. Testing writing is a hard task as writing subjective in nature. In the instruction of writing, various rubrics can be utilized for assessment depending on the aim of the grading, the type and quantity of the feedback that the scorer wants to give, the qualities and requirements of the writing task. A rubric is a scoring tool that assists evaluating the written or spoken works of learners on a continuum from excellent to poor (Schafer, 2004). According to Herman,

*ADDRESS FOR CORRESPONDENCE: Buket Demirbüken Department of Foreign Languages, Marmara University, E-mail address: buket.demirbuken@marmara.edu.tr ORCID ID: 0000-0001-7607-5381
Büşra Ulu Department of Foreign Languages, Gebze Technical University bulu@gtu.edu.tr ORCID ID:0000-0001-7607-5371

Aschbacher, and Winters (1992), rubrics have four main features: namely criteria, standards, scale, and examples. An efficient rubric includes well defined criteria that is expected from students. It also includes standards for a clear description of the proficiency levels. For those standards, there is mostly a scale to indicate to what extent they have been met. In addition, a well-defined rubric exemplifies the anticipated performance in a certain task.

The rubrics are generally classified as holistic and analytic (Hughes, 2002). Holistic rubrics include assigning "a single score to a piece of writing on the basis of overall impression of it" (Hughes, 2002, p.93). Thus, they do not allow the instructor to evaluate the constructs of writing such as organization, depth of vocabulary and syntax. They are advantageous in terms of time as it takes less to score using them. However, the reliability and validity of those types of scoring are questionable as it is dependent on subjectivity. On the other hand, the analytic scoring type assigns a separate score to each aspect of writing that is to be evaluated. It basically "includes writing components relating to the test taker's lexical, syntactic, discourse, and rhetorical competence" (Ghalib & Al-Hattami, 2015, p.227). One advantage of such rubrics is that they have higher reliability and are based on more objective judgment by the scorers. Besides, feedback given based on these rubrics is more detailed, comprehensive, constructive and beneficial for the improvement of the writing skill. Nevertheless, analytic scoring requires a lot of effort, and therefore it is time consuming. The other drawback of them is the possibility of disregarding the overall quality of a work due to inclusion of various aspects.

The use of holistic and analytic scales has been a matter of research in language education as a part of assessment. In order to shed light on this issue, many studies were conducted on the use of various rubrics and their efficiency. In the following, some pioneering studies in the literature were presented with their findings to expand our knowledge.

2. Literature Review

Assessing writing has many challenges and intricacies to be solved. There are basically two approaches to writing assessment. Holistic and Analytic Assessment: Holistic assessment is implemented for general understanding of learners' writing proficiency whereas analytic assessment is preferred to provide more detailed evaluation for instructional purposes. Each approach has its own benefits and disadvantages. First, it is crucial to decide if the assessment will serve instructional purposes or a general overview of writing proficiency. The literature offers many studies and reviews on the benefits and drawbacks of the scales in various writing contexts, some of which are given below:

In one of early studies, Bacha (2001) found out that analytic scale was more beneficial over holistic as an evaluation instrument in EFL program. In the study, the final exams of L1 Arabic Freshman students of the English language teaching program were evaluated by both holistic (Sweedler- Brown, 1993) and analytic scales (Jacobs et.al., 1981). The raters' scores were correlated by the superman correlation coefficient using SPSS. Likewise, Wiseman (2012) investigated holistic and analytic scales to assess ESL (English as a second language) learners' writings of college students. The performance of rubrics was determined by Rasch-many-faceted measurement and the raters indicated that holistic rubrics were rarely preferred by the raters with the fear of failure in creating a five-point-scale while analytic rubric was determined as a better instrument for diagnostic and placement aims. Similarly, Putri et. al. (2022) studied the efficacy of holistic and analytic scales by grading 24 eighth graders' writings. The teachers were randomly assigned to one of the scales out of holistic and analytic. The results were in line with Wiseman (2012) and Bacha (2001) as they supported the analytic scoring by claiming that analytic score was more dependable and defined the performances more superficially. In accordance with these results, Al-Ghazo and Taamneh (2021) also found that most of the Jordanian EFL teachers working in secondary school were in favor of analytic scores, and they stressed the significance of adopting a rubric for writing assessment.

As regards holistic scales, the literature seems to be hesitant to suggest holistic scales alone. Bachman (1990) emphasized the importance of validity for a good scale, which is questionable in holistic ones (Weigle 2002). Önem (2022) criticizes the holistic scales as to validity by stating that "only one score represents the whole performance" (p.13) which is confirmed by Weigle (2002). On the other hand, Barkovi (2011) conducted a study to investigate test takers' writing ability by employing two different methods (holistic vs. analytic). The results indicated that holistic scale indicated a higher inter-rater reliability while analytic scales had higher self-consistency. Likewise, Harsch and Martin (2012) studied the rating quality in grading writing papers of students at proficiency levels. The study employed both holistic and analytic scales and all raters (six in total) had some training to combine holistic scores with analytic to increase validity. The study confirmed its initial hypothesis, which was that holistic scales may pretext the discrepancies though they had higher inter-rater reliability.

More complicatedly, some other studies favor both holistic and analytic scales by differentiating the purpose of usage. Galti et. al. (2018) had a review paper on the use of holistic and analytic scales as well as pros and cons by concluding that variant aims for writing result in using different scales to measure student performance. The paper also had a suggestion for using holistic scales for beginners with a focus on content and analytic for higher levels to evaluate 'superficial features' as well as content. In a recent study, Imbler, Clark, Young and Feinauer (2023) conducted a quasi-experimental study with secondary school learners in order to find whether analytic or holistic rubric yields meaningful results. Their findings showed that the former type enabled evaluators to identify specific strengths and weaknesses of learners' writings while the latter one provided a general evaluation. Thus, they pointed out that instructors ought to evaluate the objectives behind scoring written work and choose the appropriate rubric that best aligns with those objectives. Similar to those results, He (2018) concluded that the purpose of the writing tasks plays a crucial role in determining scoring in certain contexts after a comprehensive review of literature.

There has been plenty of research about the utilization of rubric in assessment of language skills. This study aims to contribute to this area of research by exploring the use of holistic and analytic rubric in writing. Specifically, It seeks to find out the strengths and shortcomings of holistic and analytic rubrics while grading the written paragraphs of English as a foreign language learners in a university. Therefore, this study uses two holistic, TOEFL (as cited in Hughes, 2009) and ACTFL (2012), and two analytic scales Anderson's analytic scale (as cited in Hughes, 2009) and Jacobs et al. (1981).

2. Methodology

2.1. Setting

The study was conducted in a public university's English preparation class in Istanbul, Turkey. The classes are leveled according to a placement test administered and named as A1, A2 and B1 in accordance with the Common European Framework. The participants are from A2 level students. They have 25 hours of English in a week and two teachers. The teachers share the class hours, follow the same syllabus and conduct the classes cooperatively.

The writing classes are four (4) hours per week and the syllabus for writing classes are designed by bottom-up approach, from paragraph to essay writing in steps.

2.2. Participants

The participants consist of 15 students attending English preparatory school of a public university. They are all pre-intermediate (A2) students, and their age range is between 18-20. They had similar educational backgrounds and their L1 is Turkish.

The analysis of the data obtained from the participants were evaluated by three raters who are also the researchers in this study. All three raters are lecturers at preparatory schools of three different universities in Turkey. They are also PhD students at Hacettepe University Foreign Language Education department. All three raters have three to ten years of teaching experience, and each has taught writing skills at different levels many times before besides graded writing papers for various levels. In addition, all raters had some readings and discussions on the scales employed.

2.3. Data Collection

2.3.1. The task

In the regular flow of their syllabus, the participants were assigned to write an opinion essay including a thesis statement, body paragraphs -without any restrictions of the number of paragraphs-, and conclusion. The topic for the writing assignment was "The Problems of University Students", which was chosen in relation to reading and listening texts in class to iron out the possibility of impotency of generating ideas. They were asked to write a complete essay on the given topic between 250-350 words, including a title and no prompts were offered.

2.3.2. Administration of the task and collecting data

The data was collected through Google Classroom as a writing assignment. The students wrote their assignments within one class hour specifically and uploaded their finished work to Google Classroom at the end of the class. They were not allowed to use dictionaries or any course materials, and access to any kind of AI, google translate or similar applications were forbidden, and their non-usage was ensured.

2.3.3. Instruments

In order to evaluate gathered essays, two types of rubrics were chosen. The study employed two holistic (ACTFL, 2012; TOEFL, as cited in Hughes, 2009) and two analytic (Anderson's, as cited in Hughes, 2009; Jacobs et.al 1981) scales to analyze the data. These common rubrics were selected as they were the most adopted ones, and the reliability and validity of them were justified by a lot of studies. The scales were explained as follows:

Holistic Scales in the Study

TOEFL (as cited in Hughes, 2009) is one of the holistic scales used by the raters and it has several descriptors for each level of scoring from 1 to 6, each of which has five descriptions referring to the task, organization, supporting details, use of language and range of vocabulary and the descriptors are quite specific for each level while the other holistic scale which is ACTFL (2012) has 9 descriptors, gathered under three main headings, but instead of being listed as discrete points, the descriptors are provided as a paragraph for each level as *advanced high, advanced mid and advanced low; intermediate high, intermediate mid and intermediate low; novice high, novice mid and novice low*. The descriptors are more precise with some form of concrete examples.

Analytic Scales in the Study

Anderson's analytic scale (as cited in Hughes, 2009) involves 5 categories of *grammar, vocabulary, mechanics, fluency (style and ease of communication), and form (organization)* for scoring students' writing texts. Each category has six-point levels with each point having one-sentence explanations. Regarding Jacobs et al.'s (1981), as cited in Hughes, 2009) analytic score, it also involves five categories like Anderson's (as cited in Hughes, 2009). However, the names of some of the categories differ. The categories involved are *content, organization, vocabulary, language use, and mechanics.* What is strikingly different about this scale is that the ratio of points assigned to each category is different for most categories.

2.4. Data Analysis

The raters carried out online meetings for sharing the collected data and designing the analysis process. In total, three online meetings were conducted. Each rater graded the papers individually. In the grading process, two analytic and two holistic scales were used. The holistic scales used in this study were prepared by TOEFL (as cited in Hughes, 2009) and ACTFL (2012) while the analytic scales were composed by Anderson (as cited in Hughes, 2009) and Jacobs et al. (1981, as cited in Hughes, 2009). The raters carried out the grading process without sharing their own grading with the other raters. After each rater analyzed the data using each scale separately, the results from the analysis were compiled in one Excel document for comparison. The participants were each named with the letter "P" followed by a number in order to keep the participants anonymous. Their pseudonyms were represented in the tables ranging from P1 to P15. The raters were also named with the letter "R" followed by a number in order to maintain anonymity. Thus, the raters were represented with the pseudonyms R1, R2, and R3 in the data tables.

When the grading process was over and all the grades were compiled in an Excel file, the raters had an online meeting to compare and discuss the results. Based on the grading and the raters' experience of using the scales, the strengths and weaknesses of each scale were discussed, and notes were taken.

In addition, the inter-rater reliability (IRR) was calculated for each scale. The scores of the holistic scales yielded ordinal data that has grouped the proficiency level of the writings in descriptive categories, so the formula, IRR= TA/ TR*R, was used. That is, the total number of agreements is divided by the multiplication of the number of ratings by each scorer and the number of raters. In order to calculate the scorer reliability of analytic scales, inter-class correlation coefficient was used as a measure of reliability. Intraclass correlation coefficient was chosen since it can be used to calculate consistency between more than two raters (Liljequist, Elfving & Skavberg Roaldsen, 2019).

3. Results

3.1. TOEFL (as cited in Hughes, 2009) and ACTFL (2012) Holistic Scales

As shown in Table 1, the raters' grades of holistic scales developed by TOEFL (as cited in Hughes, 2009) and ACTFL (2012) were compared and presented. As the results suggest, when the grades obtained with the TOEFL scale (as cited in Hughes, 2009) were analyzed, all 3 raters graded 5 papers the same, out of 15 papers. 7 papers were graded the same by two raters, with a difference of 1 or 2 points by one rater. The remaining 3 papers were graded differently by each rater. As being experienced and trained lecturers, the raters had the same or very close grades on 12 papers and had discrepancy for 3 papers; however, this still corresponds to a quite large percentage which is 20%.

Table 1. The Grades of Raters with TOEFL (as cited in Hughes, 2009) and ACTFL (2012) Holistic Scales

Participants	TOEFL (Holistic)			ACTFL (Holistic)		
	R1	R2	R3	R1	R2	R3
P1	1	3	3	Int-low	Int-low	Int-mid
P2	2	3	4	Int-mid	Int-mid	Int-high
P3	3	4	4	Int-high	Int-high	Int-high
P4	3	4	5	Int-high	Int-high	Int-high
P5	4	5	5	Adv.-Low	Adv.-Low	Adv-low
P6	3	3	3	Int-high	Int-high	Int-mid
P7	3	3	3	Int-high	Int-high	Int-mid
P8	1	1	1	Int-low	Nov-high	Nov-high
P9	4	4	4	Int-high	Int-mid	Int-mid
P10	5	5	4	Adv-low	Int-high	Int-mid
P11	2	2	2	Int-low	Int-low	Nov-high
P12	3	4	5	Int-mid	Int-mid	Int-high
P13	3	3	2	Int-high	Int-low	Nov-high
P14	2	3	3	Int-mid	Int-mid	Int-mid
P15	4	4	5	Int-high	Int-high	Adv-low

The raters reflected that the holistic TOEFL scale (as cited in Hughes, 2009) had some strengths and weaknesses. One problem that was put forward by raters was that some descriptors were fulfilled in some students' writing while others were not present in any part of the writing. On the other hand, some writing texts had elements that fulfilled parts of the descriptors of several different levels. This makes it quite difficult to decide which level is the true level of the students. Besides, having a scale from 1 to 6 does not really provide much context for how the students perform in terms of their writing ability. This is especially because of how the descriptors are designed. For instance, "A paper in this category is adequately organized and developed" is one of the descriptors of Level 4. The descriptor does not specify what "adequately organized and developed" is. What is adequate for one rater may not be assessed as adequate by another. There is a level of vagueness in descriptors. That is, the descriptive items for each level are explained with limited measurement. Additionally, there is no place for punctuation in the scale. A writing includes many items, so some parts lack to be measured although holistic assessment enables a comfort of seeing the task as a whole and placing it to a band, which makes the scorer's job easier, and avoids the scorers from getting caught up in the small details and missing the big picture. This means that the scoring may be susceptible to subjective judgements.

When it comes to ACTFL (2012) holistic scale grades, it has differences when compared to TOEFL (as cited in Hughes, 2009) holistic scale though both aim to look at the paper as a whole and grade holistically. As being different from TOEFL (as cited in Hughes, 2009), in ACTFL (2012), the score bands are described in paragraph formats for each band and there are no discrete items that will distract the raters. As a weakness, it is possible to say that the bands are limited with writing needs. The bands under the first category NOVICE only refer to listed items, emails or notes. Paragraphs and essays are automatically meant to a higher level which may cause an evaluation gap in terms of not overlapping with genre and the content. Also, long paragraphs of descriptors make it quite difficult to evaluate writings. It is confusing and causes one to forget what to look for and how to judge written texts according to each and every one of the descriptors.

With all that ACTFL (2012) brings, the results show that all 3 raters graded 4 papers the same out of 15 papers. 9 papers were graded the same by two raters with a difference of 1 up or down score band by one rater. The remaining 2 papers were graded as different by each rater. The discrepancy rate for the ACTFL (2012) score is 13.3% which is lower when compared to the TOEFL (as cited in Hughes, 2009) holistic scale. It may be because the descriptors are defined differently, and the genre constrained the bands in ACTFL (2012) which left less space for the raters and may pave the way for the bands to be the same or close. For example, a task which is graded as 1 or 2 with TOEFL (as cited in Hughes, 2009) scale cannot be differentiated in ACTFL (2012). Likewise, 4 or 5 cannot be differentiated in ACTFL (2012) scale again. On the other hand, based on raters' experience, ACTFL (2012) befittingly ensures to look at the whole picture and to make a perfect holistic grading.

3.2. Reliability of the Scales

In the following table, the reliability between each rater is given:

Table 2. The Reliability for TOEFL Scale (as cited in Hughes, 2009)

	R1 & R2	*R1 & R3*	*R2 & R3*	*R1 & R2 & R3*
reliability=	*0,36*	*0,13*	*0,2*	*0,37*

As it can be seen in Table 2, the consistency between R1 and R2 is poor with r=0,26. Even reliability between R1 and R3 is lower. R2 and R3 have a bit more consistent scores, but despite that, reliability remains not acceptable. The overall consistency among the three scorers is 0,49. Even though it is the highest one, it is undesirable as r= 0,5 is not a good one.

Table 3. The Reliability for ACTFL Scale (2012)

	R1 & R2	*R1 & R3*	*R2 & R3*	*R1 & R2 & R3*
reliability=	*0,36*	*0,13*	*0,2*	*0,37*

As presented in Table 3, Researcher 1 and Researcher 3 had the lowest interrater reliability whereas Researcher 1 and Researcher 2 had the highest one. The overall consistency among the raters was 0,37. However, all of these values are <0,5, so they are not acceptable. It can be said that these gradings are not reliable.

3.3. Anderson (as cited in Hughes, 2009) and Jacobs et. al.'s (1981, as cited in Hughes, 2009) Analytic Scales

Table 4. The Grades of Raters with Anderson (as cited in Hughes, 2009) and Jacobs et. al.'s (1981, as cited in Hughes, 2009) Analytic Scales

Participants	Anderson (Analytic)			Jacobs et. al. (Analytic)		
	R1	R2	R3	R1	R2	R3
P1	2+2+3+3+3=13	3+2+3+2+3=13	3+4+3+4+5=19	17+13+10+18+3=61	17+14+10+17+3=61	22+15+12+17+3=69
P2	3+3+3+3+4=16	3+3+3+3+4=16	5+4+5+4+4=22	19+15+13+15+4=66	21+14+14+17+4=70	25+15+13+15+3=71
P3	3+4+4+4+4=17	4+3+3+3+4=17	5+5+5+4+4=23	20+15+15+20+3=73	22+17+15+17+4=75	25+16+14+19+5=79
P4	2+4+2+4+4=16	4+3+4+3+4=18	5+5+4+3+5=22	20+15+10+15+4=64	21+16+17+17+4=75	25+17+15+18+3=78
P5	4+4+3+4+5=20	5+5+5+4+5=24	5+5+5+4+5=24	22+17+15+18+4=76	25+18+18+21+4=86	26+17+17+20+4=84
P6	4+4+3+4+4=19	4+4+4+4+4=20	3+4+3+4+4=18	24+17+15+17+3=76	25+18+17+18+4=82	23+15+14+13+3=68
P7	3+3+3+3+3=15	4+3+3+4+3=17	5+5+4+3+4=21	19+15+15+17+3=69	22+14+14+17+4=74	23+14+14+17+4=72
P8	3+2+2+2+2=11	3+3+3+3+4=16	3+3+4+2+5=17	17+15+10+10+3=55	18+10+13+11+3=55	19+10+11+12+3=55
P9	4+4+4+5+2=19	4+4+4+4+3=19	4+4+4+4+3=19	17+13+15+20+3=68	22+13+14+17+3=69	21+13+13+16+3=66
P10	4+4+4+5+4=21	4+4+4+5+4=21	4+4+5+4+3=20	20+17+15+18+4=74	22+17+17+21+4=81	22+16+17+20+4=79
P11	2+3+2+2+3=12	2+3+3+3+3=14	2+3+3+3+4=15	15+10+10+10+2=47	17+13+10+11+3=54	18+13+13+14+3=61

In Table 4, the raters' grades of analytic scales (Anderson, as cited in Hughes, 2009, and Jacobs et. al., 1981, as cited in Hughes, 2009) were compared and presented. As the results suggest, when the grades obtained with Anderson's scale (as cited in Hughes, 2009) were analyzed, it seems that all three raters had agreement on 2 papers; 6 papers were graded the same by two raters with a difference of 1 or 3 points by one rater. The remaining 8 papers were graded as different by each rater. The highest discrepancies are observed for P4, P7 and P8 with 6 points. However, the difference for P7 is not problematic by Jacobs et al.'s (1981, as cited in Hughes, 2009) scale while the grades for P4, similarly, differ 14 points. The discrepancy between the grades is bigger by Jacobs et al.'s (1981, as cited in Hughes, 2009) scale; that is, the highest discrepancy is observed for P14 with 22 points. On the other hand, the paper of P8 had 6 points discrepancy by Anderson's scale (as cited in Hughes, 2009), all three raters agreed on the same score for this paper by Jacobs et al.'s (1981, as cited in Hughes, 2009) scale. It may be because the components of each scale differ and content has a large proportion of the overall score in Jacobs et al.'s (1981, as cited in Hughes, 2009) scale while there is no section devoted to content in Anderson's scale (as cited in Hughes, 2009). In Jacobs et al.'s (1981, as cited in Hughes, 2009) scale, the highest portion of points is dedicated to content, with possible points varying between 13 to 30, and make a difference when compared to other scales.

Apart from the distribution of points for each category, the points assigned within the categories are also in the form of point ranges (30-27 Excellent to very good, 26-22 Good to average, etc.). This type of point distribution seems to provide a basis for sound evaluation, but it is in fact quite problematic for several reasons. One reason is that the lowest score for the categories is never 0, which means that even a highly insufficient writing would get a fair number of points. The point range system is another reason why this scale is problematic. No rationale is provided as to how to decide which point to assign from the point range indicated for each level of the five categories. For instance, the highest level for the content category is "Excellent to very good", but there is no indication of how to decide on which point to assign from the 27-30-point range. However, the biggest problem emerges from the descriptors. Instead of providing a list of sentences or paragraphs, the descriptors are listed as short phrases. Most of these are vague and open to interpretation in terms of their intended meaning and there is also room for subjective evaluation. That is; by Jacobs et al. 's (1981, as cited in Hughes, 2009) scale all the raters agreed on only once (0,66 % agreement) and only for 1 paper two raters agreed with a discrepancy by the other rate (% 96.6 disagreement). The highest discrepancy is observed in Jacobs et al. 's (1981, as cited in Hughes, 2009) scale with 22 points in line with 20 points discrepancy for P14; 14 points discrepancy for P1; 13 points discrepancy for P12 & P15.

The researchers' experience with Anderson's scale (as cited in Hughes, 2009) is that having five categories provides some kind of thorough evaluation, which also gives a sense of comfort to the rater since scoring different categories separately allows the rater to judge distinct aspects of students' writing and thus, have a better understanding of the writing ability of students. On the other hand, while the scoring of writing seems to be more structured compared to the two holistic scales, there is still room for subjectivity in scoring. For example, it is difficult to differentiate the parts from each other; that is if the paper is not successful holistically, it becomes difficult to give a high mark from vocabulary and punctuation, though these parts are not problematic. The scorer may have the tendency to mark as low if the paper is not good as a whole, or in the opposite way, mark as high if the paper is good as a whole.

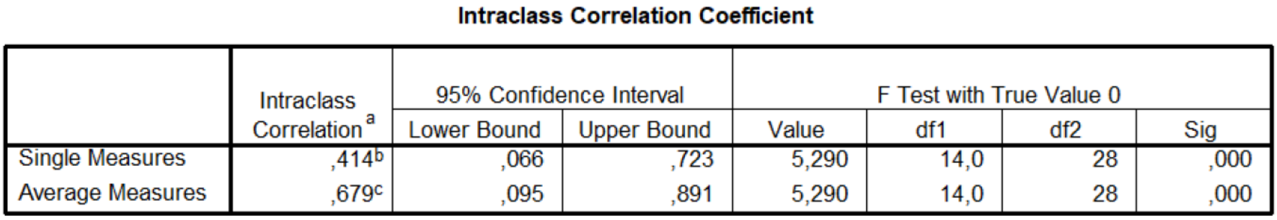
3.4. Interrater Reliability of the Scales

Table 5. Means and standard deviations of researcher for Anderson's scale (as cited in Hughes, 2009)

	Mean	Std. Deviation	N
Rater 1	16,2667	2,98727	15
Rater 2	17,8000	2,73078	15
Rater 3	20,0000	2,67261	15

Table 5 shows the means and standard deviations of 3 researchers' scores for the writings of the students. It can be seen that Researcher 3 tended to give higher grades compared to the other two raters and she has the highest mean. The means of Researchers 1 and 2 are close to each other, 16 and 17 respectively. The standard deviations (SD) of all scorers were almost the same, and nearly all scores of the researchers are scattered within two standard deviations from the mean.

Table 6. Intraclass Correlation Coefficient (ICC) for Anderson's scale (as cited in Hughes, 2009)



Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,414[b]	,066	,723	5,290	14,0	28	,000
Average Measures	,679[c]	,095	,891	5,290	14,0	28	,000

In Table 6, results of intraclass correlation are presented. For reliability, due to the number of raters, the correlation across raters is needed, so average measures results are essential. The intraclass correlation of Anderson's scale (as cited in Hughes, 2009) is 0,7. ICC value is between 0 and 1 and 1 showed higher reliability and similarity between raters' grades. In this study, r=0,7 is acceptable, and it is close to 1, so it can be interpreted that there has been a higher consistency among scorers for this scale.

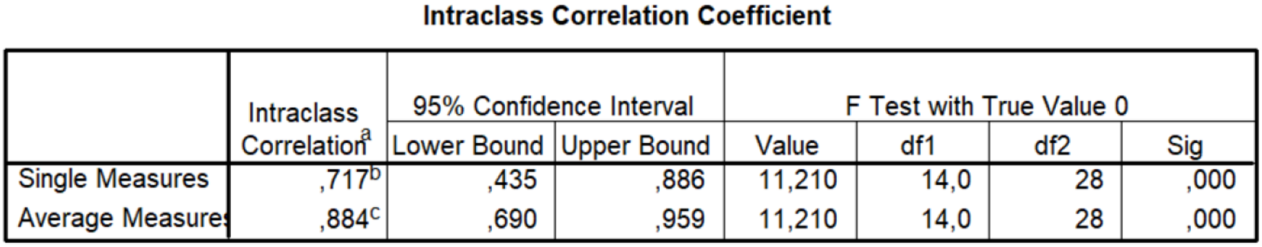
Table 7. Means and standard deviations of raters for Jacobs et. al.'s (1981, as cited in Hughes, 2009)

	Mean	Std. Deviation	N
Rater 1	66,13	8,114	15
Rater 2	70,60	9,560	15
Rater 3	71,13	9,195	15

Table 7 indicates the means and standard deviations for the Jacobs et. al.'s scale (1981, as cited in Hughes, 2009). As shown, Researcher 1 gave lower grades than other two, and had the meaning of 66. Other two scorers had the tendency to give higher grades than Researcher 1, and their means were close to each other, 70,60 and 71.13

respectively. The standard deviations of Researcher 1 were lower standard deviation, eight, while Researchers 2 and 3 had almost the same number, nine. The scores of the rater 1 were spread out within eight SD from mean while the other two raters' scores were within nine SD. deviations from the mean. Therefore, it can be understood that the scores of Researcher 1 are more close to the mean as the SD is lower.

Table 8. Intraclass Correlation Coefficient (ICC) for Anderson's scale (as cited in Hughes, 2009)



Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,717[b]	,435	,886	11,210	14,0	28	,000
Average Measure	,884[c]	,690	,959	11,210	14,0	28	,000

In Table 8, intraclass correlations are given. When average measures are interpreted, it can be said that correlation is 0,9. it shows higher reliability among raters.

4. Discussion & Conclusion

The study investigated two holistic (ACTFL, 2012; TOEFL, as cited in Hughes, 2009) and two analytic (Anderson's, as cited in Hughes, 2009; Jacobs et.al 1981) scales by grading students' writings by three raters and presents the results by comparing their scores besides sharing the raters' experience with the scale while grading. The discrepancy in comparison to raters' scores across scales allowed the researchers to interpret the reliability of the scales, which was in favor of holistic scales as regards inter-rater reliability. The findings of the current study were in line with some studies in the literature (Barkaoui, 2011; Harsch & Martin, 2012). However, this study also indicated that holistic scales have higher inter-rater reliability which may be because they offer flexibility to the raters by grading the whole picture (Hughes, 2002) by trivializing the details while analytic scales identify the problems point by point and hence limit the rater with its criterion-based nature.

In addition, the raters' experience with the scales also showed the support for holistic scales. Thus, the raters' experience with holistic scales was more convincing and comforting; on the other hand, in comparison to holistic scales, the grades were more variable in analytic scales as each component was graded separately and the total has the possibility to differ accordingly. In holistic scoring, all three raters commented on the easiness and practicality when compared to analytic scoring as it saved time and helped the raters by not enforcing them to make a detailed correction. Those findings supported claims of He (2023) who pinpointed the practicality of holistic rubrics. and however, the raters also stated that it was difficult to feel safe about holistic scales in line with Weigler (2002) and it might be difficult to give feedback to students in return. On that point, raters showed a tendency towards purposeful choice of scales (Hughes, 2002; Galti et.al., 2018, Clark et.al., 2023) as our aspirations may have a strong effect on our choice of scales.

In a study conducted by Cheng, Yang & Han (2022), raters also commented on holistic scales as "the scale is more suited to those occasions where speedy results are prioritized" (p. 12). In classroom practice where students need feedback and details, holistic scales may not be efficient. That is, when the aim is more student-oriented, analytic scales can be more efficient as both the rater (teacher) and learner have an idea about strengths and weaknesses for each component. On the other hand, the raters in this study had the highest discrepancy about Jacobs et al.'s (1981, as cited in Hughes, 2009) analytic score. It was interpreted by the raters that the point ranges are varied and too diverse for each component, which makes the job difficult for the raters. A good training and piloting is a must to ensure agreement between raters.

A number of important limitations need to be considered. Sample size of the study might interfere with the generalizability of the findings, and thus studies with a larger sample could yield more valid and reliable results. Also, for each type of rubric, two rubrics were chosen for the assessment of writings. The evaluation part with all of the selected scales was challenging and confusing for the scorers. Therefore, in order to reduce the workload of the researcher, only one rubric for each type could be selected.

The findings of the current study provide certain implications for the use and development of rubrics for writing assessment. One of the issues that emerge from this study is that both holistic and analytic rubrics have drawbacks as well as the advantages. In order to overcome these problems that emerge during assessment, one solution might be selecting an appropriate rubric that is parallel with the objectives of the scorer and the writing task. The other implication is about the possibility of lower inter or intra-rater reliability. As seen in the grades of the evaluators, reliability between raters can differ. In order to avoid such consistency issues, training for the relevant rubric that is to be used for assessment of writings is necessary.

In general, it can be concluded that both analytic and holistic scales used in this study can be useful in various circumstances. If assessment is required to be on-site with results being delivered quickly, holistic scales may prove to be more beneficial compared to analytic scales. On the other hand, analytic scales may provide more detailed and grounded feedback for the test-takers regarding their performance. Analytic scales may also allow raters to grasp a better understanding of the test-takers' writing ability based on the delivered performance when compared to holistic scales.

References

Al-Ghazov, A. & Ta'amnehi I., M. (2021). Evaluation and grading of students' writing: Holistic and analytic scoring rubrics. *Journal for the Study of English Linguistics*, *9*(1),77-93.doi:10.5296/jsel.v9i1.1906

American Council on the Teaching of Foreign Languages (ACTFL). (2012). *ACTFL proficiency guidelines.* ACTFL, INC.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, *29*(3), 371-383.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford university press.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, policy & practice*, *18*(3), 279-293.

He, L. (2018). Direct assessment of second language writing: Holistic and analytic scoring. *Westcliff International Journal of Applied Research 2*(1),37-48. doi: 10.47670/wuwijar201821LH

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, *20*(3), 281-307.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Hughes, A. (2009). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Imbler,A. C., Clark, S. K., Young, T. A.,& Feinauer, E. (2023).Teaching second-grade students to write science expository text: Does a holistic or analytic rubric provide more meaningful results? *Assessing Writing,55.*

Galti, A. M., Saidu, S., Yusuf, H., & Goni, A. A. (2018). Rating scale in writing assessment: Holistic vs. Analytical scales: A review. *International Journal of English Research*, *4*(6), 4-6.

Ghalib, T. K.., & Al-Hattami, A. (2015). Holistic versus analytic evaluation of EFL Writing: A case study. *English Language Teaching, 8*(7). doi: 10.5539/elt. v8n7p225

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation – A discussion and demonstration of basic features. *PLOS ONE, 14(7).* doi: 10.1371/journal.pone.0219854

Önem, E. E. (2022). *Holistic vs. analytic assessment of speaking in EFL-attitudes, scores and the raters*. Astana Yayınları.

Putri, E. S., & Melani, M. (2022). Holistic vs. Analytic Evaluation in Writing Test of Eighth Grade Students. *Journal of English Language Studies*, *7*(2), 157-175.

Schafer, L. (2004). *Rubric*. Retrieved February 9, 2015, from http://www.etc.edu.cn/eet/articles/rubrics/index.htm

Sweedler-Brown, C. O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing*, *2*(1), 3-17.

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

Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *International Journal of Language Testing*, *2*(1), 59-92.

Note. We thank our colleague Reyyan Zülal YÖNEY for providing support for data collection and grading for our project.