Holistic vs. Analytic Scoring between Expository and Narrative Genres: Does the Assessment Type Matter?

Sahail M. Asassfeh

Professor, Department of Curriculum & Instruction, College of Educational Sciences, The Hashemite University, Jordan
Corresponding Author: Sahail M. Asassfeh, E-mail: sabash@hu.edu.jo

ARTICLE INFORMATION	ABSTRACT
	The unprecedented emphasis on EFL students' development of their writing skills has invited researchers and EFL instructors to look for the most effective methods of teaching writing and assessing it. Within this context, two main dominant assessment types have been used by EFL writing instructors: holistic and analytic scoring with ambiguous research findings about the use of each. Moreover, ambiguity has also been surrounding variability in raters' scoring across genre types. This study aims at uncovering the difference between using the two scoring schemes across two genres in EFL writing: expository and narrative. Two texts each representing a genre type from 10th grade EFL writers' compositions were presented to a sample of 90 in-service EFL teachers for holistic scoring. Two months later, the same texts were presented to the same teachers for analytic scoring in order to compare between the grades assigned in each round for arriving at results. Results suggest significant differences between the scores obtained according to the grading method. Also, narrative essays received higher scores in both assessment types. These results are discussed, and recommendations are derived.

1. Introduction

The common framework of looking at the educational process typically involves three major phases: planning, implementation, and assessment or evaluation. The last component of these involves whatever activities teachers perform to use as a source of feedback for improving the teaching/learning process (Black & William, 1998).

According to Brown and Abeywickrama (2010) in their seminal work, *Language Assessment*, one dimension in language assessment that was raised in the 1970s and still has an impact amounts to the nature of language. Accordingly, two approaches have emerged: discrete-point and integrative testing. Whereas the first relies on the assumption that language can be viewed as a sum of divisible components (i.e., listening, speaking, reading, and writing) and multiple units of language (phonology, graphology, morphology, lexicon, syntax, and discourse), the latter argues that language abilities are indivisible, hence should not be assessed individually. According to the latter assessment type, integrative testing, the best to assess language are cloze tests and dictations, for these require integration of all major language abilities at a time. It is for this reason that the skill addressed in the current study is writing, for it encompasses many other skills within.

There is an agreement among EFL researchers that decisions regarding which type of assessment is more effective are not to be made arbitrarily. The general impression is that analytic scales are more effective if what is sought are precision and valuable feedback whereas propriety is assigned for holistic scoring if feedback is not at the heart of the assessment purposed (Thomas, 2020).

As commonly used today, assessment can refer to two different activities: (a) the mere gathering of information and (b) the use of that information for improvement (evaluation), and her lies the border between the two inextricably linked terms of assessment and evaluation (Astin, 1991; Bachman, 2004). That is, saying that a student deserves a grade of X is assessment, but whether this means pass or fail at one level and why the grade was not higher are two concerns that fall under evaluation—but not assessment.

Thomas (2020, pp. 240-1) summarized five considerations to be kept in mind when designing or selecting an assessment scale. These involve what exactly is going to be assessed and for what purpose, how much time is required--to ensure practicality, the validity and reliability of the scale, and whether using the scale requires prior training.

In one very recent study, Vögelin, Jansen, Keller, Machts, and Möller (2019) addressed the difference between the outcomes of holistic and analytic scoring on a sample of four argumentative writing pieces with variability in lexical diversity and sophistication graded by 37 pre-service teachers in Switzerland. It appeared that lexical diversity and sophistication had a positive impact on the grades holistically assigned. And even though variability was limited to the lexical dimension, analytic scoring components (grammar and essay overall frame) were also influenced positively.

The current study has ample significance for EFL theory and practice. One target is the exposition of the differences and their implications between adopting a holistic compared to analytic scoring of EFL writing. Likewise, an exposition of the impact of genre type (expository vs. narrative) on EFL teachers' scoring of EFL students' writing is of paramount significance. At a practical level, EFL assessors are in bad need for research-based practice so that the results of their assessment would be trustworthy, valid, and reliable. Theory-wise, no final say has been reached regarding the effectiveness of the two (holistic vs. analytic) scoring scheme, a research concern to which this study is likely to contribute. No firm answers have also been reported regarding the impact of the genre type on EFL teachers' scoring practice and results. At a practical level, EFL teachers and supervisors as well as testing agencies can find the outcomes of this study beneficial. Assessment, if carefully handled, can be a good start for the effective teaching that educators seek. This study finding should come as a response.

2. Literature Review
2.1 Analytic scoring
This is a method of scoring a student's composition, which depends on a marking scheme that has been carefully drawn up by an examiner or a body of examiners. The procedure consists of an attempt to separate the assessment criteria.

Analytic scoring entails ample focus on various text features for assigning a grade. These might include such text components and traits/ criteria as content, organization, vocabulary, grammar, and mechanics, cohesion, coherence, etc.

The use of separately delineated scales may guard against the possibility that raters will collapse categories during the rating process, since they must produce a separate rating for each scale (Astika, 1993). In this method of marking, each separate criterion in the mark scheme is awarded a separate mark and the final mark is the composition of these individual estimates.

Several advantages have been reported in the literature (cf. Hosseini & Mowlaie, 2016) for analytic scoring. For example, it provides students with a clear idea about how the total score is assigned and based on what divisions. It also enhances the interrater reliability and uncovers any mismatch in students' development of various subskills. Moreover, it heightens raters' sensitivity to and awareness of text aspects that could have been ignored otherwise. It also functions as an important ingredient of formative assessment whose outcomes can be used as a guide for determining where students' challenges lie for further effort and concentration.

These advantages do not come at no expense, for obviously they consume much time and decontextualize text aspects from each other and their surrounding context, minimizing the raters' sense of the overall text as one integrated unit.

2.2 Holistic scoring
Holistic scoring is a method of scoring students' composition in which one or more evaluators read the paper without marking anything, and then rate the paper as a whole, assigning single scores based on total impression of composition as a whole text or discourse. The holistic scoring method is based on the theory that a whole piece of writing is greater than sum of its parts (Bacha, 2001).

Hosseini and Mowlaie (2016) investigated the impact of analytic and holistic scoring on the writing development of Iranian EFL learners. Each of two intermediate-level experimental groups received a different type of scoring (holistic versus analytic). Results revealed that whereas both scoring methods had a statistically significant impact on students' writing quality measured by the post-test, the group receiving analytic scoring did significantly higher on the post-test.

In their recent study in the Turkish context, Han and Huang (2017) compared between holistic and analytic scoring in terms of the impact on scoring variability and reliability in EFL institutional writing. A writing sample of 36 argumentative essays were graded by 10 raters holistically and then analytically. Results suggested that effective training can yield grades comparable in

reliability to those assigned by analytic grading. The results suggested that holistic scoring was favored by raters. And among the grading criteria, content was the most decisive one in holistic scoring whereas all criteria (e.g., grammar, content, or organization) had influence on the outcomes of analytic scoring.

Huang and Han (2013) conducted a study on nine randomly selected EFL undergraduate students' argumentative essays representing high, medium, and low-quality levels at a Turkish university. Each was graded both holistically and analytically by five raters in an attempt to investigate: (a) any variability associated with the scoring type; (b) the sources of score variability between the scoring plans, and (c) any difference in the reliability of the scoring between the scoring plans. Results revealed consistently higher scores for analytical compared to holistic scores on seven essays, higher variability among the scores resulting from holistic scoring, and lower reliability for holistic scoring, with overall results favoring analytic over holistic scoring for its appropriateness and effectiveness.

Despite the abundance of studies that have addressed EFL assessment, there is consensus among specialists that assessment overall has not yet received the due attention. Moreover, digging deep at the level of specific issues related to assessment has rarely been a characteristic of EFL assessment research. For example, there are no decisive results pertinent to the difference between holistic and analytic scoring. More specifically, it is not clear whether text genre matters in raters' scores assigned to a written text, which this study aims to uncover through answering the two following main questions:

1. Are there any statistically significant differences ($\propto$ = .05) in EFL teachers' scores assigned to a writing piece based on the assessment type (holistic vs. analytic?
2. Are there any statistically significant differences ($\propto$ = .05) in EFL teachers' scores assigned to a writing piece based on the type of writing genre (expository vs. narrative)?

3. Methodology

This study follows a quantitative and descriptive design. It was conducted on a convenient sample of 75 EFL teachers who were willing to participate in the study. Nonetheless, among those who participated in the first grading round (holistic), only 48 were successfully reached in the second round, and they provided their assessment results via email to their supervisors. Thus, the number of participants whose responses filtered into the statistical analysis for the purpose of the study was 48. All had an experience in teaching the 10th grade or were teaching it during the time of conducting the study. One consideration related to the selection of the 10th grade relates to text complexity and length. EFL teachers teaching the 11th and 12th grade were intentionally avoided because of their duties in preparing students for the Secondary School Exit Exam (*tawjihi*). Intentional avoidance of these two grades was intended to avoid contamination of results that could have resulted from grading under pressure. The highest grade left, therefore, was the 10th. The researcher sought the help of EFL supervisors who knew EFL teachers well and knew the grades they were teaching. As for the participants' academic qualifications, some held a Higher Diploma or a Master's degree, whereas the majority of them were BA holders. This variability is thought to have no impact on the findings of the study since both texts were introduced to each and the same participant twice.

3.1 Instrument of the study

Two texts each representing a genre were developed for the purpose of this study. One represented expository and the other represented narrative writing. The texts were constructed to represent the writing of 10th graders. Semantic and linguistic complexity was also considered for equivalence. Essay length determined by the number of words was considered and the topic of both essays was also intentionally the same.

3.2 Validity and reliability of the Instruments

A sample of each text was presented to a panel of three experienced EFL teachers and two EFL supervisors. Informed about the purpose of the study, each referee was requested to judge the following: (a) unity of the topic in both essays; (b) appropriateness of the text to resemble the writing of 10th grades whose writing is very common to the participants of the study; and (c) lexical-item equivalence in terms of complexity. Particularly in a study of this type, consistency in grading was ensured by the fact that whether there is lenience in grading or not, the two text types were graded by the same participant twice. It was decided that holistic grading should come first, and paper-based grading was a requirement so that during the second round of grading (analytic), it would be unlikely that a copy of the text could have been kept and a period of two-month time span between EFL teachers' encounter with the text was secured. Four criteria were included in the rubric for analytic scoring: grammar, vocabulary, cohesion, and task completion. These criteria were adopted from a version of grading writing from IELTS Exam (rubric attached). Each criterion was assigned 10 marks. The highest possible score, therefore, was 40.

3.3 Data Collection

The data for this study was collected with help of EFL supervisors who were familiar with the participants of the study and the grades they taught. Essays were presented to teachers and collected by their supervisors during their work hours each inside his/her school in the first round (holistic grading). However, for the circumstances of the emergence of COVID-19, the second round of grading (analytic) was provided by EFL teachers via e-mail to their supervisors and forwarded to the researcher. The number of responses on each of the two scoring rounds was 48.

3.4 Statistical Analyses

For the purpose of answering the research questions, the quantitative data for the current study was fed into and analyzed using SPSS 21 to arrive at the results.

4. Results

As noted earlier, the current study aimed at investigating the difference between two assessment types (holistic vs. analytic). It also aimed at examining whether there are differences in scores assigned for two equivalent essays attributed to the type of text genre. To answer the first question, which addresses the differences between the two assessment types, paired sample *t*-test was used. The two pairs compared were: (a) holistic vs. analytic scoring with expository texts and (b) holistic vs. analytic scoring with narrative texts. The results (Table 1) indicate that there were statistically significant differences ($\propto$= .05) between holistic and analytic score regardless of the genre type.

Table 1. *T*-test results of mean score comparisons between the two types of assessment with each genre

		Paired Differences							
				Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
		Mean	Std. Dev.		Lower	Upper			
Expo.	Holistic vs Analytic	5.60	2.77	.40	4.80	6.407	14.039	47	.000
Nar.	Holistic vs Narrative	6.19	2.63	.38	5.43	6.951	16.294	47	.000

Descriptive statistics results suggest that the mean score for holistic grading of expository essays was higher than the value for analytic scoring (*M*= 27.33, *SD*= 1.85). Likewise, with narrative essays the mean score for holistic scoring *M*= 35.48, *SD*= 1.75) was higher than its counterpart for analytic scoring (*M*=29.29, *SD*= 1.75).

Whereas the first research question addressed differences associated with the assessment type, the second targeted differences by genre type. Although holistic grading results in a single value for the entire essay, analytic scoring results in a value for each essay accompanied by a value for each criterion within the rubric assigned for analytic grading. As mentioned earlier under instrumentation, the four components of analytic scoring included grammar, vocabulary, cohesion, and task completion. Results (Table 2) suggest a statistically significant difference associated with the following mean score comparisons:

1. Holistic grading by essay type
2. Analytic grading by essay type
3. Analytic grading of vocabulary by essay type.
4. Analytic grading of cohesion by essay type.

However, there were no statistically significant differences in the analytic grading of both genre types associated with grammar or task completion.

Table 2. *T*-test results of mean score comparisons between the two types of assessment and analytic assessment criteria between genres

		M	D	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Holistic	E vs N	-2.54	.47	36	-3.26	-1.83	-7.14	47	.000
Analytic	(E vs N)	-1.96	.97	43	-2.82	-1.10	-4.57	47	.000

Analytic	Grammar (E vs N)	.375	.86	?7	-.17	.916	1.39	47	.170
Analytic	Vocab (E vs N)	-1.06	.21	I7	-1.41	-.71	-6.08	47	.000
Analytic	Cohesion (E vs. N)	-1.04	.30	I9	-1.42	-.66	-5.54	47	.000
Analytic	Task (E vs N)	-.23	.13	I6	-.56	.10	-1.40	47	.168

- E= Expository, N= Narrative

The mean score assigned holistically for narrative essays was relatively higher than that assigned for expository essays. This extends to the results of analytic scoring as well. Descriptive results presented in Table 3 show which genre type yielded higher mean scores. At the level of rubric criteria for analytic scoring, relatively higher scores were assigned for narratives consistently, except for grammar.

Assessment Type/component	Min	Max	Sum	Mean	SD
Holistic-Narrative	32.00	39.00	1703.00	35.48	1.75
Holistic -Expository	28.00	37.00	1581.00	32.94	2.17
Analytic- Narrative	25.00	34.00	1406.00	29.29	1.99
Analytic Expository	22.00	31.00	1312.00	27.33	1.85
Analytic Narrative- Vocabulary	6.00	9.00	375.00	7.81	.790
Analytic-Narrative- Cohesion	5.00	9.00	362.00	7.54	.90
Analytic Expository Grammar	5.00	10.00	347.00	7.23	1.02
Analytic Narrative Task Completion	6.00	9.00	340.00	7.08	.94
Analytic narrative Grammar	4.00	10.00	329.00	6.85	1.43
Analytic Expository – task completion	5.00	9.00	329.00	6.85	1.01
Analytic Expository - Vocabulary	5.00	8.00	324.00	6.75	.84
Analytic- Expository Cohesion	5.00	9.00	312.00	6.50	.97

5. Discussion

This study aimed at answering two main questions. The first concerns any statistically significant differences in EFL teachers' scores assigned to a writing piece based on the assessment type (holistic vs. analytic), whereas the second addresses any statistically significant differences in EFL teachers' scores assigned to a writing piece based on the type of writing genre (expository vs. narrative). The results are confirmatory to the abundance of significant differences between the scores obtained according to the grading method. Higher scores were associated with holistic scoring regardless of the text genre. That is, an essay that describes an event in sequence yields higher scores compared to one that is closer to the scientific type of text. Several justifications are likely to stand behind this finding. One plausible justification might relate to the cognitive load assigned in the case of holistic grading that is commonly referred to as impressionistic. In other words, holistic grading might be less capable of arousing the grader's sense of focus on specific standards or criteria. Graders, accordingly, look at the overall picture of the essay without depicting details pertinent to particular standards for grading. Analytic scoring, on the other hand, demands keeping in mind a set of factors to look for while grading. The more the criteria, the higher the concentration required. More concentration also seems to raise more concerns about any text regardless of genre. One more possible justification relates to the purpose behind assessment. It is commonly the case that there is a difference between a score on which important decisions are to be made in contrast to scoring for an overall picture reflecting the writer's command on the writing components and standards. Accountability also might play an important role in determining how deep a rater's critical eye should be at play. Clearly, important decisions require higher level of accountability. This explains why most rigorous international exams require using a rubric with a clear set of criteria. Precision, therefore, seems to be associated with analytic in contrast to holistic scoring. This finding goes in line with previous findings suggesting that analytic scoring triggers a more critical eye and demands only *justifiable* results. This is not the case in holistic grading where it is sufficient to give an overall result impressionistically without a need for rigorous justifications. The level of concentration required on the grader's behalf is of paramount significance in analytic in contrast to holistic grading. Educators in general and writing scorers, therefore, seem to be in need to adopt analytic scoring when grading a text based on grading where an important decision will be made. Moreover, educators should be urged to use analytic scoring in the case of teaching academic in contrast to literary writing as has been indicated by the findings of Hosseini and Mowlaie (2016). The reciprocal relationship between teaching and assessment invites writing instructors to help students become more aware of the components of a text and text characteristics. Grading should also be fair; when analytic scoring is going to be adopted, students should be aware of the standards based on which their writing is going to be judged accordingly from the very beginning. It seems unfair to train students on holistic scoring and apply analytic scoring, which inevitably requires training on analyzing text elements and characteristics Han and Huang (2017). Careful emphasis on the use of formal lexical items and grammatically correct linguistic sentence structures in addition to text aspects such as cohesion and task

completion becomes a necessity in writing instruction. Writing, in this sense, is a science that has rules to follow and serious concerns to keep in the writer's mind.

Results pertinent to the second research question that addresses grading across genres suggest almost consistent findings associated with the genre type. Higher grades consistently accompanied narrative texts compared to expository ones. This finding might be attributed to several factors. One important dimension amounts to amusement. As commonly known, people often read stories for entertainment, where ideas are presented with the purpose of creating amusement in the reader. This goes in line with the findings of Han and Huang (2017). This is not the case when compared to expository writing that has the aim of informing the reader about a topic s/he is expected to be in need to "learn" about. It is likely that a narrative invites more attention to the plot and sequence of events, which does not require an analytic eye. Especially when the events are well presented in a logical sequence, graders might tend to assign a relatively high grade regardless of text qualities and characteristics. Whereas this might be the case, it should be avoided when grading for language learners who are "learning to write" and working on developing their writing abilities. Another factor that might well be at play relates to the type of genre in terms of text characteristics. Expository writing is quite often characterized by density and richness of ideas. An expository text invites reading on a table rather than a coach, which reflects the level of alertness required on the reader's behalf. This alertness is more likely to help the reader in depicting any area of weakness in a given text.

6. Conclusion

The findings of this study confirm that holistic scoring is more likely to yield higher grades compared to analytic scoring. Narrative texts are more likely to yield higher grades compared to expository ones. The take-home lesson based on these findings yielded rigorous, accurate, reliable, and precise assessment of writing.

The current study has ample significance for EFL theory and practice. One target is the exposition of the differences and their implications between adopting a holistic compared to analytic scoring of EFL writing. Likewise, an exposition of the impact of genre type (expository vs. narrative) on EFL teachers' scoring of EFL students' writing is of paramount significance. At a practical level, EFL assessors are in bad need for research-based practice so that the results of their assessment would be trustworthy, valid, and reliable. Theory-wise, no final say has been reached regarding the effectiveness of the two (holistic vs. analytic) scoring scheme, a research concern to which this study is likely to contribute. No form answers have also been reported regarding the impact of the genre type on EFL teachers' scoring practice and results. At a practical level, EFL teachers and supervisors as well as testing agencies can find the outcomes of this study beneficial. Assessment, if carefully handled, can be a good start for the effective teaching that educators seek. This study finding should come as a response.

Analytic scoring is to be a priority, a target that is hard to accomplish without sufficient training. Content in both types of scoring is a decisive factor in shaping students' grades on written tasks. And if EFL teachers are to help EFL writers develop their texts, teacher training on using grading rubrics is inevitable. Finally, this research is not without limitations. The number of the participants in future research is recommended to be higher for more generalizability.

References

[1] Astika, G. G. (1993). Analytical assessment of foreign student's writing. *RELC Journal, 24*, 61-72.
[2] Bachman, L. F. (2004). Statistical analyses for language assessment. Cambridge: Cambridge University Press.
[3] Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80* (*2*), 139-148.
[4] Brown, H. D. & Abeywickrama, P. (2010). Language Assessment: Principles and classroom practices. White Plains, NY: Pearson Education.
[5] Huang, J., Han, T. (2013). Holistic or analytic – A dilemma for professors to score EFL essays? *Leadership and Policy Quarterly*, *2*(1), 1-18.
[6] Hosseini, M & Bahram Mowlaie, B. (2016). Effect of holistic vs. analytic assessment on improving Iranian intermediate EFL learners' writing skill. *Journal of Language and Translation, 6*(11), 31-41.
[7] Han, T. & Huang, J. (2017). Examining the impact of scoring methods on the institutional EFL writing assessment: A Turkish perspective. *PASAA*, *53*, 112-147.
[8] Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL. *Assessing Writing*, *39*, 50-63.