Research Journal of English Language and Literature (RJELAL)
A Peer Reviewed (Refereed) International Journal
Impact Factor 6.8992 (ICI) http://www.rjelal.com;
Email:editorrjelal@gmail.com; ISSN:2395-2636 (P); 2321-3108(O)

Vol.14.Issue 1. 2026 (Jan-March)

REVIEW ARTICLE

INTERNATIONAL STANDARD SERIAL NUMBER INDIA
2395-2636 (Print):2321-3108 (online)

Evaluating Writing Tasks and Marking Approaches in Language Testing: A Critical Analysis

Nguyen Thi Le Phi
Lecturer of English, Ho Chi Minh City University of Natural Resources and Environment, Vietnam
Email: ntlphi@hcmunre.edu.vn

DOI: 10.33329/rjelal.14.1.453

Article info
Article Received: 01/03/2026
Article Accepted: 24/03/2026
Published online: 31/03/2026

Abstract

This paper critically evaluates a writing task designed for teenage English as a Foreign Language (EFL) learners and compares two marking approaches used to assess their written performance. Drawing on key concepts in language testing such as reliability, validity, authenticity, and washback, the discussion examines the strengths and weaknesses of the task and its scoring methods. The analysis highlights the importance of clear task specification, appropriate assessment criteria, and inter-rater reliability in ensuring fair and meaningful evaluation as well as emphasizes how well-designed assessment practices can enhance both measurement quality and positive backwash in language learning. In addition, relevant literature review in language testing is reviewed to strengthen the theoretical foundation of the discussion. The paper concludes by emphasizing the pedagogical implications of well-designed writing assessments and analytic scoring procedures.

Key words: Language testing, assessment, evaluation, reliability, validity, authenticity, washback.

1. INTRODUCTION

Language testing plays a crucial role in language education because it provides evidence of learners' language ability and helps educators evaluate the effectiveness of teaching practices. In both classroom and large-scale assessment contexts, tests serve multiple purposes, including measuring language proficiency, assessing achievement, diagnosing learning difficulties, and informing instructional decisions.

Designing effective language tests requires careful consideration of fundamental principles in language testing theory. According to Bachman & Palmer (1996), useful language tests should demonstrate several key qualities, including reliability, validity, authenticity, interactiveness, impact, and practicality. Among these qualities, reliability and validity are often

considered the most fundamental measurement properties, while authenticity and backwash are particularly important in educational contexts where assessment directly influences teaching and learning practices.

Writing assessment presents particular challenges for language testers and classroom teachers. Writing ability is widely viewed a complex construct involving multiple linguistic and cognitive components such as grammar, vocabulary, discourse organization, and coherence (Bachman & Palmer, 1996; Weigle, 2002). Because writing assessment frequently relies on human judgment, scoring reliability may be affected if assessment criteria are unclear or inconsistently applied (Hughes, 2003). Consequently, both the design of writing tasks and the scoring procedures used to evaluate them play a critical role in determining the quality of writing assessment.

Previous research has extensively examined scoring reliability and rubric design in writing assessment (Song et al., 2025). Analytic scoring rubrics and rater training have been shown to enhance scoring reliability by standardizing evaluation criteria and reducing rater variability (Weigle, 2002). In addition, washback research has highlighted the impact of assessment practices on teaching and learning processes (Cheng et al., 2015; Qin & Jia, 2025). Despite these developments, relatively fewer studies have examined how classroom-based writing tasks and teacher-developed scoring procedures align with established principles of language testing theory.

This gap is particularly relevant in many EFL classroom contexts where teachers design assessment tasks independently and apply locally developed scoring methods. Without systematic evaluation of these practices, classroom assessments may inadvertently compromise reliability, validity, or fairness.

Therefore, the present study aims to evaluate a writing task and two marking approaches used in a classroom-based writing assessment for teenage EFL learners. Drawing on key concepts in language testing theory—reliability, validity, authenticity, and washback—the study seeks to identify potential strengths and limitations in both task design and scoring procedures.

The study addresses the following research questions:

1. To what extent does the writing task meet the principles of reliability, validity, and authenticity in language testing?
2. How do the two marking approaches differ in terms of scoring reliability and transparency?
3. What implications do these findings have for improving classroom-based writing assessment practices?

2. LITERATURE REVIEW

2.1 Reliability in Writing Assessment

Discussing about a reliable test, Heaton (1989) indicates that reliability refers to the consistency of a test as a measurement instrument, ensuring consistency and accuracy in results. A reliable assessment should produce stable results across different testing occasions, raters, and scoring procedures. In writing assessment, reliability is particularly challenging because evaluation often involves subjective human judgment. In short, in order to be reliable, a test must be consistent in its measurements or the more similar the scores would have been, the more reliable the test is said to be (Hughes, 2003).

Previous studies have demonstrated that the use of analytic scoring rubrics and standardized scoring procedures can improve scoring consistency among raters. For example, Song et al. (2025) found that structured scoring rubrics significantly increased agreement between human raters in large-scale writing assessments.

2.2 Validity in Language Testing

According to Hughes (2003), a test is said to be valid if it measures accurately what it is intended to measure. We create language tests in order to measure such essentially theoretical constructs as "reading ability", "fluency in speaking", "control of grammar", …. For this reason, in recent years the term construct validity has been increasingly used to refer to the general, overarching notion of validity.

- Content validity: The test would have content validity only if it included a proper sample of the relevant structures. Just what are the relevant structures will depend, of course, upon the purpose of the test. In order to judge whether or not the test has content validity, we need a specification of the skills or structures, etc. that it is meant to cover. A comparison of test specification and test content is the basis for judgments as to content validity.
- Criterion – related validity: there are essentially two kinds of criterion – related validity: concurrent validity and predictive validity
- Face validity: a test is said to have face validity if it looks as if it measures what it is supposed to measure. According to Heaton (1989), if a test looks right to other testers, teachers, and testees, it can have face validity. That's why we, testers should show a test to colleagues and friends. Only if the test is examined by other people can some of the absurdities and ambiguities then be discovered. Moreover, the students' motivation is maintained if a test has good face validity.
- Validity in scoring: it is worth pointing out that if a test is to have validity, not only the items but also the way in which the responses are scored must be valid. It is no use having excellent items if they are scored invalidly.

2.3 Authenticity in Writing Tasks

Authenticity refers to the extent to which test tasks resemble real-world language use situations (Bachman & Palmer, 1996). We need to be able to demonstrate that performance on language tests corresponds to language use in specific domains other than the language test itself. It is this correspondence that is at the heart of authenticity. We define *authenticity* as the degree of correspondence of the characteristics of a given language test task to the features of a target language use task. Authenticity is important because it has effects on test takers' perceptions of the test and, hence, on their performance.

However, authenticity also depends on learners' familiarity with task topics. If students lack personal experience with a given topic, their performance may reflect background knowledge rather than writing ability. Consequently, task designers must consider learners' sociocultural and experiential backgrounds when selecting writing prompts.

2.4 Washback Effects of Writing Assessment

Washback is the effect that tests have on learning and teaching (Hughes, 2003). It is now seen as a part of the impact a test may have on learners and teachers, on educational systems in general, and on society at large.

Also discussing the effect that tests have on learning and teaching, Phuong et al., (2025) indicates that assessment methods can shape classroom instruction, learning strategies, and student motivation. Research has shown that analytic rubrics can promote positive backwash by helping students understand the components of effective writing. However, other studies suggest that teachers may interpret rubrics differently, which can lead to variation in classroom assessment practices.

In recent bibliometric research, Qin & Jia (2025) emphasizes that washback has become a major focus in applied linguistics research over the past three decades, particularly in

Asian educational contexts. These findings highlight the importance of aligning assessment design with pedagogical goals in language education.

3. METHODOLOGY

3.1 Research Design

This study adopts an evaluative research design to examine the effectiveness of a classroom-based writing task and two marking approaches. The evaluation is grounded on theoretical principles of language testing, particularly reliability, validity, authenticity, and washback.

The study focuses on a classroom assessment context in which a writing task was administered to teenage EFL learners and evaluated using two different marking procedures. The analysis aims to identify strengths and limitations in both the task design and the scoring methods.

3.2 Participants

The participants consisted of ninety teenage learners of English as a Foreign Language (EFL) studying in a classroom context. The learners had pre-intermediate and intermediate-level English proficiency and were enrolled in a language course designed to develop fundamental writing skills.

3.3 Writing task

The writing task was administered as part of a classroom-based assessment activity intended to evaluate students' basic writing ability.

"Think of a time when you were travelling somewhere and the journey was very long. Discuss your experience."

3.4 Marking Procedures

Marking Approach 1: Student scripts were divided among several teachers, who assigned scores based on multiple language features such as spelling, punctuation, grammar, control, and argument.

Marking Approach 2: In the second approach, two teachers evaluated each script using a combination of holistic and analytic scoring. Raters first formed an overall impression of the text and then assessed specific language features including grammar, vocabulary, punctuation, coherence, and sentence structure. When scoring differences occurred, the raters discussed their judgments and reached a mutually agreed score.

4. DISCUSSION

4.1 Writing task

The above writing task examined in relation to key assessment principles such as clarity of instructions, genre specification, and topic accessibility revealed several issues related to the design of writing task. First, there is a topic for the learners to write about but they certainly wonder themselves if they have to write this in thirty or forty minutes and in how many words they have to express their ideas. The teacher wants them to discuss the experience of a long journey which may include many things to share while he or she gives no limit of the words included. Moreover, the learners may confuse about the genre they are going to apply. Whether a letter writing to someone is a good idea or a narrative telling someone about the experience of a long journey is what the teacher wants them to share. This question surely leads a confusion to the learners who always need a clear instruction for each question.

Surely, the writing task lacked explicit instructions regarding genre and expected length. Without such guidance, students may interpret the task in different ways. For instance, some students might produce a narrative describing personal experiences, while others might write a descriptive account or a reflective essay. These variations introduce additional sources of variability that are unrelated to writing ability, thereby potentially reducing the reliability of the assessment. Because the task did not specify expectations regarding structure

or genre, evaluating these different responses using the same scoring criteria becomes challenging. Clear instructions certainly help the teacher find it easy to assess the work. So this test tends to be unreliable since it doesn't have one of the features of a reliable test which contains items or tasks that are unambiguous to the test takers.

Second, the topic of a "very long journey" may not be equally accessible to all learners. Some students may have had extensive travel experiences, while others may have had limited opportunities to travel. So the learners who have never had a chance for a long journey may lose motivation with such an irrelevant task. As a result, students with richer experiential backgrounds may be able to produce more detailed and engaging narratives, even if their linguistic ability is similar to that of other students. As being consistent with previous research about language testing, this paper strongly recommends that the above writing task should be focused on the word "a time" and "somewhere"; the teacher should indicate clearly "a time" here as "the summer holiday" which all of the learners have experienced and "somewhere" here as "a countryside, a farm house or a village, a beach…" which are really close to the teenage learners. This issue may affect both task authenticity and fairness.

4.2 Marking approach

Given the two marking approaches, this paper supposes that there are two groups of teenage learners in order to have some detailed discussion.

The first group consists of individuals aged 11 to 15 whose English level is at intermediate low and they only can meet some limited writing skills. With this aspect, marking approach 1 seems to be valid because it focuses on what it is intended to measure which are some basic features to mark on a writing task including spelling, punctuation, grammar,… However, with this group, it is hoped that the aspect of "control" and "argument" should not be the language features to be assessed. Moreover, In Marking Approach 1, scripts were distributed among different teachers without systematic moderation procedures, which may lead to the fact that each teacher interpreted the scoring criteria independently, variation in scoring judgments may occur. For example, one teacher might prioritize grammatical accuracy, while another might give greater weight to content development. This practice may result in an invalid scoring.

The other learners are in the second group of the ages between 15 to 19 who certainly have a better language competence, so five components in marking approach 1 used to assess their writing skills are not enough for the teachers to measure the learners' ability. Marking approach 2 with more specific language features including grammar, vocabulary, punctuation, coherence and sentence structure are more persuasive for this group. Moreover, Marking Approach 2 involved two raters evaluating each script using both holistic and analytic perspectives. This process allowed raters to compare their judgments and resolve discrepancies through discussion. Such moderation procedures are widely recommended in writing assessment because they help ensure greater scoring consistency and transparency.

It is strongly recommended that the teachers should have a suitable specified criterion because if the analytic method is employed, it is necessary for them to pay attention to the levels of learners. If the learners in the group of the ages from 11 to 15, the teachers should be more interested in spelling, grammar, vocabulary rather than cohesion or argument and the other group of learners who are in upper level should be scored in a different way. The use of multiple scoring in marking approach 2 reveals that it is a reliable scoring because "the more scores for each candidate, the more reliable should be the total score" (Hughes, 2003, p. 94).

Moreover, in marking approach 2, the teachers compared their scoring. Multiple scoring allows discrepancies to be identified and discussed, which can help reduce individual bias and improve scoring consistency (Song et al., 2025). If there were differences they reviewed the script in relation to the marking criteria at a mutually agreed upon score. I think that there have been a lot of unreliable markers because of many different reasons one of which is "the failure to agree with colleagues on the relative merits of a student's composition." (Heaton, 1989, p. 144)

4.3 Pedagogical Implications

The findings of this study have several implications for teachers designing writing assessments for teenage learners. Writing tasks should provide clear instructions regarding genre, length, and expectations. Task topics should also be familiar and relevant to students' experiences in order to encourage meaningful responses. Finally, scoring procedures should be transparent and supported by analytic rubrics and collaborative evaluation practices. Such approaches can improve the reliability of writing assessment while providing constructive feedback that supports students' writing development.

5. Conclusion

The findings of this study provide several insights into the design and evaluation of classroom writing assessments for teenage learners. The analysis suggests that both task design and scoring procedures play an important role in ensuring reliable and valid assessment outcomes. First, the clarity of task instructions appears to be a key factor affecting students' performance. Second, the findings highlight the importance of selecting topics that are accessible to teenage students. As Bachman & Palmer (1996) note, tasks should reflect authentic communication while remaining appropriate to learners' backgrounds. When topics require experiences that some students may not possess, the assessment may inadvertently measure background knowledge rather than language ability. Finally, the comparison of marking approaches indicates that scoring procedures influence the consistency of writing evaluation.

This study evaluated a classroom-based writing task and two marking approaches using principles from language testing theory. The analysis revealed several limitations in the design of the writing task, including unclear instructions and limited topic accessibility, which may affect reliability and validity.

The comparison of scoring procedures demonstrated that the second marking approach, which combines analytic and holistic scoring with rater moderation, provides more consistent and transparent evaluation outcomes.

Overall, the study underscores the importance of carefully designed writing tasks and systematic scoring procedures in classroom language assessment.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching, 48*(4), 436–470.

Heaton, J. B. (1989). *Writing English language tests*. Longman.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Phuong, H. Y., Le, T. T., & Pham, T. T. (2025). Assessment for learning or assessment for scoring? Washback of analytic rubrics in Vietnamese EFL classrooms. *Language Testing in Asia, 15*, Article 32. https://doi.org/10.1186/s40468-025-00371-y

Qin, X., & Jia, Q. (2025). Three decades of research on washback (1993–2023): A bibliometric study. *Language Testing in Asia, 15*, Article 23. https://doi.org/10.1186/s40468-025-00357-w

Song, D., Lee, W. C., & Jiao, H. (2025). Exploring LLM autoscoring reliability in large-scale writing assessments using generalizability theory. *arXiv preprint*. https://arxiv.org/abs/2507.19980

Weigle, S. C. (2002). Assessing writing. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997