Journal of Language Teaching, 1(1), 1-5, 2021
https://doi.org/10.54475/jlt.2021.001

Review

A comparative study of analytical assessment and holistic assessment in English writing tests

Hao Wu [1, 2]

[1] **The University of Oxford, Oxford, United Kingdom**
[2] **Naval Medical University, Shanghai, China**

Received: September 25, 2021 / Accepted: October 3, 2021 / Published Online: October 5, 2021

Abstract

The choice of assessment method is particularly important in English writing testing and assessment. Analytical assessment and holistic assessment are two common methods in English writing assessment. The choice of assessment method usually depends on the content and focus of the test. This paper provides a review of previous empirical studies and compares the two assessment methods in terms of reliability, construct validity, practicability, and impact, followed by suggestions in the selection, design, and operation of the assessment methods.

Keywords analytical assessment; holistic assessment; English writing; writing assessment

1. Introduction

As one of the L2 productive skills, writing ability is of vital importance in global communication. Meanwhile, the teaching of writing is also an increasing role in second language education. Wherever the instruction of L2 writing is given close focus, the assessment of L2 writing gains equal attention. This article discusses and contrasts two prevalent writing assessment methods: analytic scales and holistic scales. It starts by giving accounts of the scales. Then, adopting the Bachman and Palmer (1996) framework, this article compares and contrasts the two scales from the perspective of reliability, construct validity, practicability, and impact. Finally, based on theoretical and empirical evidence in the comparison, this article offers practical advice for assessment methods in language classrooms.

Analytic scales describe criteria of different aspects that can be used to evaluate a test taker's achievement in particular skills (e.g., vocabulary, grammatical accuracy, organization, coherence, and register). Scores are separately awarded to each scale, and the overall estimate of an assigned task is usually the sum of each scale. Analytic scales, therefore, provides a more exhaustive report about a test takers' merits and weaknesses in specific writing areas (Bacha, 2001; Weigle, 2002). In particular, a test administrator can assign a different weight or value to a criterion, depending on their predefined requirements for the task. Appendix 1 provides a typical kind of analytic scale for a writing test. A four-point scale (1: "poor" to 4: "good") was used to assess a *script*'s achievements in five equally weighted criteria: originality of content, organization, vocabulary, grammar, and cohesion. A script refers to the written text assessed by a rater. The final score is the sum of the five criteria.

In contrast to analytic scales, holistic scales assign only one score to a script (e.g., a letter grade, a number, a percentage, or other ordinal scales). In a typical holistic writing assessment session, raters give a score that reflects their overall impression of test takers' performance in the task. Raters are trained rigorously so that their judgment adheres to a *rubric*, which is a tool with descriptors and criteria for raters to evaluate test takers' performance on a scale (Campbell, Melenyzer, Nettles, & Wyman Jr, 1999). In the rater training, benchmark scripts selected from students' authentic responses are often provided to exemplify the criteria of each level (Bacha, 2001; Weigle, 2002). As an illustration, Appendix 1 shows how the Test of English as a Foreign Language Internet-Based Test (TOEFL iBT) assesses the independent writing using a 0-5 scale (ETS, 2019).

2. Contrasts between analytic scales and holistic scales

Given the accounts of the two scales, neither scales are the optimal choice in every writing assessment. In other words, both have a place in authentic practice since the two scales have their unique advantages and unavoidable weaknesses in different scenarios and teaching settings. The Bachman and Palmer (1996) framework of test usefulness helps provide insights into their core qualities. This section discusses and contrasts four core qualities: reliability, construct validity, practicality, and impact. Each sub-heading follows the routine of giving accounts of a quality, summarizing typical empirical articles, and providing critical analysis.

2.1. Reliability

Reliability estimators in a rubric describe the consistency across raters' judgment on test takers' performance. High reliability means that raters can generate similar decisions and give similar scores to test takers (Ghalib & Al-Hattami, 2015). Two kinds of reliability are usually discussed in research articles: inter-rater reliability, one descriptor measuring different raters' rating consistency across the same script (Nakamura, 2004), and intra-rater reliability, another descriptor estimated by having the same rater rate the same script more than once (Cohen, 2017).

Analytic scales are usually assumed to have higher reliability than holistic scales (Hamp-Lyons, 1991; Huot, 1996; Weigle, 2002). As a typical example, Ghalib and Al-Hattami (2015) provided rigorous statistical evidence to support this standpoint. Thirty Yemeni EFL undergraduate students majoring in English participated in Advanced Writing Skills in their seventh semester. They were the top 30 students with the highest GPA in their cohorts. Three raters were invited to the study and given a two-hour training session. The writing task asked the participants to write a 250-word descriptive essay in the given time. The raters were first asked to rate the 30 scripts using the holistic scale, and a month later, they were asked to rate the same scripts using the analytic scale to guarantee an independent judgment.

The ANOVA demonstrated that the differences between the three raters were insignificant when they used the analytic scale ($F_{(2, 87)} = 0.373$, $p = 0.690$) but significant when they used the holistic scale ($F_{(2, 87)} = 4.833$, $p < .05$). These findings suggest that analytic scales yield more consistent and reliable scoring results. Moreover, the intra-class correlation coefficient (ICC) indicated that the three raters' intra-rater reliability was higher when they used the analytic scale. The ICC under the analytic scale was .958 with a 95% confidence interval and .797 under the holistic scale. Higher ICC means higher intra-rater reliability. However, the study did not provide sufficient data to prove that the inter-rater reliability of the analytic scales is higher than that of the holistic scales.

Overall, this study illustrates reliable and detailed results showing that analytic scales are advantageous in providing more consistent and reliable assessments than holistic scales. However, there are limitations related to the sampling and the generalizability of the results. First of all, the sample size is too small. Only 30 students participated in the study. It can be problematic to generalize those findings to the larger EFL learner groups. Moreover, the 30 students were those who obtained the highest GPA in the same department. Although convenience sampling guaranteed the homogeneity of the sample, the reliability of the study was negatively affected because the authors overlooked learners with lower English proficiency.

Despite the problems, this study is consistent with East and Young (2007), Jonsson and Svingby (2007), and Nakamura (2004), who confirmed the merits of the reliability of analytic scales. Notably, Knoch (2009) trained ten raters to rate 100 scripts and found that analytic scales have higher inter-rater reliability than holistic scales in an EFL EAP context, which complemented the lack of the inter-rater reliability measurement in Ghalib and Al-Hattami's study. Moreover, in a similar study, Zhang, Xiao, and Luo (2015) selected 300 scripts from 5,000 Chinese EFL students by stratified sampling and obtained consistent conclusions with Ghalib and Al-Hattami. All the evidence boosts the confidence in recognizing analytic scales as a more reliable scoring method than holistic scales in the college-level writing assessment. With these in mind, future research in more learning contexts (e.g., primary schools and secondary schools) is urgently encouraged to make a broader generalization.

2.2. Construct validity

Validity in writing assessment uses test results to answer the question "Have the rating scales measured what the test administrator wanted to measure?" Among various aspects of validity, construct validity has been given the most attention. Assessment scales have high construct validity when their scoring results can distinguish the representativeness of writing skills and performance (Bacha, 2001; Jonsson & Svingby, 2007; Weigle, 2002).

Analytic scales are considered to have higher construct validity than holistic ones. The ESL Composition Profile provided by Jacobs et al. (1981) is a sophisticated analytic scale with high construct validity (see Appendix 3). The five components of the scale are clearly illustrated: content, organization, vocabulary, language, and mechanics, with each one expounding well-defined rating descriptors (e.g., "excellent to very good" and "good to average," corresponding

explanations, and numerical scales (e.g., "30-27" and "26-22"). Moreover, benchmark scripts of each were also provided for raters. In this case, significant differences in test takers' individual writing skills can be effectively informed by the scores in each construct.

In contrast, capturing test takers' performance in a single score, holistic scales cannot provide construct-relevant assessment, leading to relatively low construct validity. After all, holistic scales do not allow criteria to be scored separately. For example, one cannot easily distinguish what makes a script scored 4 worse than one scored 5 in a TOEFL independent writing test. Is that coherence, organization, or vocabulary? For the same script, if rater A unconsciously weights organization more, and rater B on vocabulary more, and the script happens to have a well-developed organization, it is possible the script is scored higher by rater A. As such, the low construct validity would negatively impact the inter-rater reliability. Even worse, in a study using think-aloud protocols with authentic raters' metalanguages, Li and He (2015) found that when employing holistic scales in a China's national English test, raters concentrated only on a limited set of criteria or even on the criteria not listed in the scales (e.g., handwriting and length), thus severely reducing the validity of the rating. Although this situation can be mitigated by rater training, providing benchmarks, and additional guidelines (Lumley, 2005), those circumstances are still unavoidable, as holistic scales per se are inferior in construct validity.

2.3. Practicability

The practicability of scales is measured by the consumption of time and money in the whole rating process. In most cases, a longer rating time means higher costs. Research has unsurprisingly reached a consensus that holistic scales spend far less time than analytic scales in rating the same script because raters using analytic scales need more time to give their scores in multiple scales (Davies et al., 1999; Hunter, Jones, & Randhawa, 1996). Using Jacobs et al.'s (1981) scale is undoubtedly more time-consuming due to its complexity and unequal weights in the scales. Some research provided quantitative evidence to support this view. Bauer (1981) claimed that analytic scales take twice as long as holistic scales in rater training and four times in grading. Identically, Zhang et al. (2015) reported that rating the 300 scripts under the analytic scale took the same batch of 14 raters up to 8.5 days, whereas 1.5 days under the holistic scale. Apparently, compared with analytic scales, holistic scales are more cost-effective.

2.4. Impact

Impact of scales refers to the effect or influence that the assessment has on a test taker. By this definition, Weigle (2002) argues that analytic scales are more helpful in providing diagnostic information, rater training, teaching design, and placement. On the contrary, holistic scales provide a single score, which may mask an imbalanced writing ability development. Empirical studies suggest that students would benefit more from analytic scales because they were informed what to improve. Teachers can also gain a whole picture of students' weaknesses to adjust pedagogy and make promotional decisions (Bacha, 2001). On rater training, research suggests that analytic scales facilitate the training process, as inexperienced raters can more rapidly comprehend and employ the criteria than holistic ones (Weir, 1990).

3. Summary and Recommendation

This brief review elaborates on the differences of reliability, construct validity, practicability, and impact between holistic scales and analytic scales. It can be assumed that whereas holistic scales should be praised for their cost efficiency in the rating practice, analytic scales are advantageous in reliability, construct validity, and returning instructive advice to language classrooms. Still, test administrators must consider the best combination of the qualities before deciding which one to use in their situation.

As scale choosing is never clear-cut, responsible recommendations for L2 writing assessment can be made only when the overall situation is taken into careful consideration. This section serves to give advice about scale selection, scale design, and the scoring process in language classrooms.

Scale selection should consider the aims of the test. Due to their practicability, holistic scales are a better choice for large-scale assessment or urgent needs of placements that have to be completed in limited time with limited recourses. Nevertheless, if a writing test is used for research purposes or to provide diagnostic information to teachers and students, analytic scales are a better option. Research has shown that analytic scales promote rating transparency (Jonsson & Svingby, 2007), rater reliability, teachers' reflective thoughts on instructional practices (Beeth et al., 1999; Luft, 1999; Waltman, Kahn, & Koency, 1998), and students' self-perception and evaluation (Schamber & Mahoney, 2006).

After the scale selection, test administrators can start the scale design. A good scale, like Jacobs et al. (1981), should give a clear, explicit, interpretable definition of the skills, scoring levels, weights (if applicable), and how the scores are reported. Those are not only for the raters but also for all stakeholders involved in the test, so clarification is always appreciated. Moreover, another critical factor to consider is the focus of the assessment. If a teacher aims to assess the

acquisition of a particular language skill, such as the past simple in a lower-intermediate level non-academic class, grammatical accuracy should be given more weights instead of organization and coherence. On this occasion, analytic scales are advantageous because they can intuitively increase the weights of grammar in the rubrics and decrease others (e.g., 50% grammar and 50% others) and give detailed feedback to students.

Once the scale is established, some procedures are needed to assure reliability and validity in the rating process. First, benchmarks are likely the most helpful tool to increase raters' consistency and agreement. Benchmarks act as the anchors exemplifying how scores are given under a scale. A set of three to ten scripts is an ideal number for raters' frequent reference during the rating. When a rater is dealing with problematic scripts, such as borderline cases, benchmarks can provide suggestions (Weigle, 2002). Studies have also confirmed the critical role of benchmarks in the rating practice and how heavily raters relied on them (Denner, Salzman, & Harris, 2002; Popp, Ryan, Thompson, & Behrens, 2003). Therefore, benchmark selection should be given extra care. Second, rater training is necessary, especially in large-scale assessment. Empirical studies provided reliable evidence to support that rater reliability can be improved by rater training, although they cautioned that variations cannot be eliminated entirely (Stuhlmann, Daniel, Dellinger, Kenton, & Powers, 1999; Weigle, 1999). Remarkably, Rezaei and Lovorn (2010) discussed the relationship between benchmarks and rater training and pointed out that the reliability and validity of scales can be guaranteed only when rater training is carefully implemented.

4. Conclusion

This article set out to give accounts and typical examples of analytic scales and holistic scales in L2 writing assessment. The above-mentioned empirical studies have confirmed that by giving a score to each criterion in a rubric, analytic scales have higher reliability, construct validity, and give back more diagnostic information to teachers and students. In contrast, holistic scales are cost-effective and timesaving in that it assigns only one score to a script. Still, continued efforts are needed to determine whether those conclusions can be generalized to language classrooms beyond the college level. Taken together, both scales have their place in the L2 writing assessment. The article holds that scale selection should cater to teachers' spatial and temporal needs. Meanwhile, test administrators should make straightforward scale designs and descriptors and conduct comprehensive rater training with benchmarks to guarantee the reliability and validity of the assessment.

References

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*(3), 371-383. https://doi.org/10.1016/S0346-251X(01)00025-2

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

Bauer, B. A. (1981). A study of the reliabilities and the cost-efficiencies of three methods of assessment for writing ability.

Beeth, M. E., Cross, L., Pearl, C., Pirro, J., Yagnesak, K., & Kennedy, J. (1999). A continuum for assessing science process knowledge in Grades K-6.

Campbell, D. M., Melenyzer, B. J., Nettles, D. H., & Wyman Jr, R. M. (1999). Portfolio and performance assessment in teacher education.

Cohen, Y. (2017). Estimating the intra-rater reliability of essay raters. *Frontiers in Education, 2*(49). https://doi.org/10.3389/feduc.2017.00049

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing* (Vol. 7). Cambridge: Cambridge University Press.

Denner, P. R., Salzman, S. A., & Harris, L. B. (2002). Teacher work Sample assessment: An accountability method that moves beyond teacher testing to the impact of teacher performance on student learning.

East, M., & Young, D. (2007). *Scoring L2 writing samples: Exploring the relative effectiveness of two different diagnostic methods* (Vol. 13): Applied Linguistics Association of New Zealand.

ETS. (2019, April 5, 2021). TOEFL iBT Test Writing Rubrics. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf

Ghalib, T. K., & Al-Hattami, A. A. (2015). Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching, 8*(7), 225-236.

Hamp-Lyons, L. (1991). Pre-text: Task-related influences on the writer. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.

Hunter, D. M., Jones, R. M., & Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation, 11*(2), 61.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*, 549-566. https://doi.org/10.2307/358601

Jacobs, H., Zinkgraf, S., Wormuth, D., Hearfiel, V., & Hughey, J. (1981). Testing ESL composition: A practical approach.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304. https://doi.org/10.1177/0265532208101008

Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly, 12*(2), 178-212. https://doi.org/10.1080/15434303.2015.1011738

Luft, J. A. (1999). Rubrics: Design and use in science teacher education. *Journal of Science Teacher Education, 10*(2), 107-121. https://doi.org/10.1023/A:1009471931127

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*: P. Lang.

Nakamura, Y. (2004). *A comparison of holistic and analytic scoring methods in the assessment of writing.* Paper presented at the 3rd annual JALT Pan-SIG Conference.

Popp, S. E. O., Ryan, J. M., Thompson, M. S., & Behrens, J. T. (2003). Operationalizing the rubric: The effect of benchmark selection on the assessed quality of writing.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39. https://doi.org/10.1016/j.asw.2010.01.003

Schamber, J. F., & Mahoney, S. L. (2006). Assessing and improving the quality of group critical thinking exhibited in the final projects of collaborative learning groups. *The Journal of General Education*, 103-137. https://doi.org/10.2307/jgeneeduc.55.2.0103

Stuhlmann, J., Daniel, C., Dellinger, A., Kenton, R., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Reading Psychology, 20*(2), 107-127. https://doi.org/10.1080/027027199278439

Waltman, K., Kahn, A., & Koency, G. (1998). Alternative approaches to scoring: The effects of using different scoring methods on the validity of scores from a performance assessment.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

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

Weir, C. J. (1990). *Communicative Language Testing*. London: Prentice Hall.

Zhang, B., Xiao, Y., & Luo, J. (2015). Rater reliability and score discrepancy under holistic and analytic scoring of second language writing. *Language Testing in Asia, 5*(1), 1-9. https://doi.org/10.1186/s40468-015-0014-4