VALIDATING AN ADAPTED B1-LEVEL ANALYTIC RATING RUBRIC FOR EVALUATING INDONESIAN STUDENTS' EFL WRITING PERFORMANCE

T-1

Listiani Listiani*, Marianne Nikolov, Ágnes Hódi*****

**University of Szeged, Doctoral School of Education*
***University of Pécs, Institute of English Studies*
****University of Szeged, Institute of Applied Pedagogy*

Keywords: interrater reliability; adapted analytic rating scale; EFL writing

This paper presents how a newly developed rating scale worked in a study assessing Indonesian students' writing performance in English as a foreign language. We examined the interrater reliability of an adapted B1-level analytic writing rubric we use for assessing twenty-three students' written texts in a larger study. The project investigates the impact of teacher audio feedback on undergraduate freshmen writing performance in an EFL writing course in Indonesia in the academic year 2022/2023. The rating scale employed was adapted from Euroexam International (2019) and a study by Thi and Nikolov (2022). We adjusted the semantic descriptors of the task achievement (TA) criterion to align them to the writing task. The rest of the criteria (coherence and cohesion/CC, grammatical range and accuracy/GRA, and lexical range and accuracy/LRA) remained unchanged. Each criterion had scores on four point scales (0 to 3). Two experts who are full professors in TEFL (Teaching English as a Foreign Language) and a researcher of literacy assessment and development, with over ten years of experience as lecturers and researchers, were invited to assess ten coded texts with various qualities. For each criterion, they gave the texts scores between 0 and 3 on a provided scoring form based on the rating scale. The texts were taken from students' written works participating in the larger study. Interrater reliability was measured using the intraclass correlation coefficient (ICC, Gliner et al., 2017, p. 186). The ICC is calculated from the scores given by the two independent raters. The 95% confidence intervals of the ICC estimates were calculated using SPSS® statistical package version 25 based on a 2-way mixed effects model, consistency, and absolute agreement. The results of the ICC calculation indicated that the ICC mean was acceptable for consistency ($r = .99$), for absolute agreement ($r = .98$), and each criterion with $r = .90$ in TA, $r = .97$ in CC, $r = .98$) in GRA, and $r = 1.00$ in LRA ($p < .05$). The interpretation of ICC values demonstrated excellent reliability. In conclusion, the B1-level rating scale was reliable and can be used to score the participants' written texts (pre- and post-tests), measuring the impact of the teacher audio feedback on Indonesian students' EFL writing performance. This study suggests that a reliable rating scale is significant to be employed for rating written works to get consistent and stable scoring. The validated scale worked well with these two raters on this task. However, if new raters are involved and a new task is used, data should be collected on how the reliability might change.