

Beyond Boundaries: Advancing Language Assessment Through Tech-Driven Training

Jacob Petersen^{1,*} and Daniel Newbury²

¹Iwate University, Japan

²Ibaraki Prefectural University of Health Sciences, Japan

Abstract

This research explores the efficacy of a tech-driven training program designed for language assessors in a Japanese university setting. Traditional rater training methods often result in inconsistencies, leading to subjective grading. To address these issues, an online training course was developed. Preliminary findings indicate that collaborative training, supplemented by interactive assessments and mixed media, can enhance raters' ability to apply grading criteria consistently. The study highlights the importance of fostering collaborative learning to improve assessment accuracy.

Keywords: Rater training, Collaborative learning, Interactive assessments

Introduction

Amidst the development stage of an English interactional assessment for Japanese university students, it became apparent that training raters to apply the rating criteria consistently and as the test designers intended would require more than a cursory approach. Rating inconsistencies and subsequent discussions among raters early in the test design process revealed several factors that may affect the rating process, including rater effects (Myford & Wolfe, 2004), subjectivity (Matthews, 2021), and the amount of detail needed in criteria definitions (Knoch, 2009). To compound the issue, finding enough qualified raters in rural Japan, where the research is being conducted, hindered the project's progress. To address these issues, the researchers considered a learning management

*Corresponding Author E-mail: jacobp@iwate-u.ac.jp

system, Moodle, that was already being used in another part of a research project. Moodle supports autonomous learning, various collaboration tools, assessment with instantaneous feedback, and seamless navigation of learning units.

Consistency and accuracy are integral to the success of rater-graded speaking assessments. For raters, a comprehensive understanding of the grading criteria is a prerequisite first step; however, bringing raters to the level of being able to apply these criteria in a fair way is the central imperative for an assessor training program. Research on the use of technology in language education is often student centered, with limited exploration of its role in educator training, including training raters to assess student performances. Assessor training may rely on text-heavy materials and a calibration session, after which raters are put onto the job of assessing (Ali et al., 2013). This approach may result in inadequately trained raters who struggle to consistently apply the assessment criteria.

Methodology

The first version of this training program was an ad hoc system that attempted to train five language instructors on the assessment criteria. These raters-in-training provided feedback through a survey and a discussion session to aid in future development. Through reviewing this information, some issues came to light. First, more concrete examples of performances at different levels were requested. Lack of collaborative opportunities to clarify their understanding was also cited as a source of discontent.

The current small-scale research is based on our second attempt at creating a training program, based on the feedback and observations from the first attempt. Bloom's Taxonomy was chosen as a guiding framework by which to ensure that the raters-in-training could learn the precepts of rating as well as how to apply those to the actual rating process (Huitt, 2011).

This second program was designed with a mixed-method research approach that entailed collecting data from a cohort of five assessors using a training course designed in Moodle. This course contained a module for each rating criterion, as well as an introduction to the test. After course completion, the raters-in-training completed a survey, and semi-structured interviews analyzed thematically were conducted. The research question was: Does collaborative online training, aided by interactive assessments and mixed media, prepare assessors for the task of assessing?

Results and Discussion

Initial results show that the new program may have improved. The raters-in-training specifically addressed the ability to collaborate when in

doubt on rating criteria, thus leading to the beginnings of self-regulated learning. This is promising because it shows that trainees are able to oversee their own learning (Sitzmann & Ely, 2011).

Central to the approach in this training program was a focus on chunking complex information into digestible segments for the raters-in-training. It is thought that a significantly lesser cognitive load facilitated a deeper understanding of the rating criteria. Guided-learning techniques, such as voice-over explanations and video-integrated quizzes, directed assessors' attention to criteria succinctly, enhancing the raters-in-training's ability to apply the criteria in a way the training designers intend. Interactions among raters-in-training bolstered confidence through the ability to attain second opinions, fostering a supportive learning environment. Process maps completed by the assessors-in-training offer a visualization of how they are implementing the rating criteria, enabling training designers to monitor and mediate assessors' understanding and application of the criteria (Papajohn, 2002). Collectively, these tools may support the transformation of the assessors' knowledge of the rating criteria into the ability to apply it, a transformation from lower-order to

Table 1. Survey results and comments

	Scores (Scale of 1–5, 5 Being the Best)	Themes & Raters-In-Training Own Comments
Course version 2 design compared to version 1	5 out of 5 raters scored the new design as 3 or better	Ease of use • “Easy to navigate.” • “Easy to understand.”
Comparing previous course tasks to the new course tasks	5 out of 5 raters scored the new course tasks as 3 or better	Ease of use • “Every task was useful in understanding the rubric.”
Mapping task (Papajohn, 2002)	1 rater scored it as a 1, the other 4 raters scored it as a 3, 4, or 5	Issues/shortcomings • “It is an interesting idea, but needed more scaffolding in the course, I was a little confused.” -By the rater who gave a score of 1
Discussion tasks	5 out of 5 raters scored as 3 or better	Task impressions • “Very useful, it was nice to discuss the rubric with other participants and hear their thoughts.”
Video examples	5 out of 5 raters scored as 4 or better	Task impressions • “I wanted more.” • “I think they helped make the rubric clearer.”

higher-order thinking skills (Krathwohl, 2002). This training approach supported through technology may help the accuracy and consistency of grading through making the experience more informative, interactive, and collaborative. Table 1 shows comments on the tasks and scores on the new course design.

Based on the preliminary results of this study, the training program is advantageous for raters as it incorporates interactive learning experiences and supports a comprehensive understanding of the material. Moreover, the program benefits test developers by enabling monitoring of raters' progress and evaluation of their foundational understanding of the rating criteria. Future directions in this research will be to scale it to a larger rater-in-training pool. Furthermore, the current researchers will continue modifications to each of the activities upon which this promising online training method is built.

References

- Ali, H. I. H., & Al Ajmi, A. A. S. (2013). Towards quality assessment in an EFL programme. *English Language Teaching*, 6(10), 132–148. <https://doi.org/10.5539/elt.v6n10p132>
- Huitt, W. (2011). Bloom et al.'s taxonomy of the cognitive domain. *Educational Psychology Interactive*, 22, 1–4.
- 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>
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
- Matthews, J. (2021). Using inter-rater discourse to trace the origins of disagreement: Towards collective reflective practice in L2 assessment. *RELC Journal*, 54(1), 99–113. <https://doi.org/10.1177/0033688220977373>
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 460–517). Maple Grove, Minnesota: JAM Press
- Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly*, 36(2), 219–233. <https://doi.org/10.2307/3588333>
- Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin*, 137(3), 421. <https://doi.org/10.1037/a0022777>