Chen, Shi
Northern Arizona University

Rater Variability in a Paired Speaking Task: A Mixed-methods Approach

This mixed-methods study intends to examine rater variability in a paired speaking task. Four female EFL raters from China participated in the study. All of them had at least one-year teaching experience. This study used a concurrent mixed-method approach. The raters graded 15 paired speaking tasks, which included 30 test takers. The speaking task was retrieved from an achievement test in Program in Intensive English (PIE) at a major university in the United States. Test takers' performances were rated based on a 4-point scale which included 13 subcategories. The preliminary results of the factor analysis provide some validity evidence of the revised rating scale. The mixed-effect analysis confirms that the raters have a significant effect on participants' ratings. Additionally, raters have demonstrated more variance in the task completion and interaction model than the linguistic features model. The results of the study are able to inform rater training, quality control and rating scale design.

Keywords: paired-speaking task, interactional competence, factor analysis, mixed-effects model

References

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*, 341–366.

Davis, L. E. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience* (Unpublished doctoral dissertation). University of Hawai'i at Manoa, Hawaii.

Ducasse, A. M. (2010). *Interaction in paired oral proficiency assessment in Spanish: Rater and candidate input into evidence based scale development and construct definition*. Frankfurt: Peter Lang.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155-185.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, 9(3), 270-292.

Gorsuch, R. L. (1983). Factor analysis. 2nd. *Hillsdale, NJ: LEA*.

Han, Q. (2016). Rater cognition in L2 speaking assessment: a review of the literature. *Working Papers in Applied Linguistics and TESOL*. doi: 10.7916/D8MS45DH

Hashemnezhad, H. (2015). Qualitative content analysis research: A review article. *Journal of ELT and Applied Linguistics*, *3*(1).

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, *33*(7), 14-26.

Richards, K. (2003). *Qualitative inquiry in TESOL*. Springer.

McNamara, T. (1996). *Measuring second language performance*. London & New York: Longman.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133- 142.

Nakatsuhara, F. (2004). *An investigation into conversational styles in paired speaking tests*. Unpublished master's dissertation, University of Essex, Essex, United Kingdom.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143-154.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? Assessing Writing, 12, 26–43.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, 8(2), 127-145.

Knoch, U., Fairbairn, J., & Huisman, A. (2016). An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test.