Research Summary:


UTSA student: Mingxia Zhi

Research on second language (L2) learners’ speech production generally utilizes human judgments to determine speech production outcomes. However, little attention has been paid to the reliability and validity of the measure (i.e., human ratings). To fill in this research gap, Huang and Jun (2014) focuses on the rater differences in the assessment of foreign accents. The authors used a quasi-experimental design to examine the perceptions of raters from different linguistic backgrounds and with varying degrees of teaching experience. The three groups of raters included in the study were inexperienced native English speaker (NES) raters, experienced NES raters, and advanced nonnative English speaker (NNES) raters. All raters evaluated 64 speeches produced by both NESs and NNES who arrived in the US at different ages, and who thus varied in their English language proficiency.

The authors found that raters’ evaluation of the relative degrees of foreign accents among speakers is reliable regardless of raters’ own native language background, suggesting that advanced NNESs can be just as competent as NESs if the assessment tasks only involve rank ordering the speakers by their degrees of foreign accents. However, the results revealed differences in the severity of ratings as a function of raters’ background. Overall, the experienced NES and advanced NNES rater groups exhibited the same degree of severity during their rating process, but the inexperienced NES raters were stricter in evaluating the speeches and assigned lower scores to the speakers than the other two rater groups. Furthermore, experienced NES raters were also better at identifying NESs than the other two groups.

The authors concluded that first language backgrounds and language experience have a significant influence on rater’s rating scores and ability to distinguish NESs from near-native early L2 learners. The researchers hence prompted L2 researchers to consider the measurement issues in L2 acquisition research. To improve the validity of L2 acquisition research and the related theories, raters’ backgrounds should be taken into consideration if raters are to be recruited for research purposes. One might misinterpret the data and draw incorrect inferences and conclusions if the measurement issue is overlooked.

To conclude, the current study raises questions about “the general reliability and validity of instruments used in measuring L2 outcomes”, and draws attention to the possible methodological improvement in L2 acquisition research (Huang & Jun, 2014). Results from the current study are in line with the conclusion by Winke, Gass and Myford (2013) that rater’s familiarity of test takers’ accents may influence their ratings. Testing programs, therefore, should consider raters’ backgrounds as a potential bias that might impact the validity of the speaking assessment. Having a second or even a third bilingual rater from different native language backgrounds, or incorporating different types of accents in rater training programs, are some of the possible solutions to improving the validity of speaking assessments.
Reference