
Reviewed by Vanessa Mari, BBL Doctoral Student

With the ever increasing use of standardized tests, research on the reliability and validity of assessments has become critically important. This study explored the rater effect in speaking assessments, using speech data from the Test of English as a Foreign Language (TOEFL). Specifically, it examined the reliability and validity of using human judgments in speaking assessments by investigating the effects of raters’ background (i.e., familiarity with accents and language teaching experience) on their judgements of non-native English speakers’ oral language proficiency.

The participants in this study included three groups of raters who varied in their familiarity with Chinese accents and their experience in teaching English as a second/foreign language. The researchers used a mixed methods approach, investigating both raters’ numerical ratings of the speech files and their written, reflective comments collected from a survey. All rater participants evaluated both the holistic (overall proficiency) and the analytical dimensions (foreign accent, grammar/vocabulary, and content) of each speech sample. This intricate methodology allows for a systematic examination on whether raters’ background has a significant effect on their rating decisions.

The results of the numerical ratings indicated no statistical differences between the three groups of raters under unguided rating context. It was also found that raters attended to similar
evaluative features while rating the speech samples. However, raters who were familiar with foreign accents and had teaching experience distinguished themselves from the other two groups because they "were more adept at separating analytical dimensions to arrive at independent ratings for each dimension, and they were better at verbalizing their thinking using linguistic terminology” (p. 782). This finding demonstrates that there is a difference in raters’ evaluative process as a function of their background, which would be an area for further research and for educators/testing companies to consider when employing human judgments to determine test-takers’ language proficiency outcomes. In addition, the qualitative portion of this study revealed that participants self reported that their experiences did have an influence on the way they rated. It would be interesting to further evaluate the relationships between raters’ self reported beliefs and their ratings.

In sum, even though there were no statistically significant differences between the three rater groups’ numerical ratings, participants in the study believed and reported that their teaching experience and familiarity with foreign accents did enhance their listening comprehension and error detection, as well as influenced their rating decisions. This finding of a potential rater effect/bias is an important one to consider for both researchers and practitioners because of the popularity of using human judgments in speaking assessments.