Title of Project:
Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of oral proficiency, accentedness, and comprehensibility

Researcher:
Ching-Ni Hsieh
Michigan State University
chsieh@ets.org

Research Supervisor:
Dr. Paula Winke

Summary of Research Findings

This dissertation project contributes to the field of language assessment—the TIRF’s research priority for which the project was funded—in three main aspects. First of all, this study examines the quality of human ratings in oral performance assessment within a high-stakes testing context. Such an investigation aligns with TIRF’s goal to ensure that English as a second/foreign language is tested in a manner that is demonstrably fair and effective. Secondly, this study includes linguistically naïve undergraduate raters, who are the main stakeholders in the international teaching assistants (ITAs) testing situation. Results of the study help identify comprehension difficulties that undergraduate students experience with ITAs and shed light on ITA testing. Thirdly, this study employs a mixed-method design in data collection and analysis, a research design TIRF encourages. This methodology helps provide a deeper understanding of rater effects on performance assessment as different aspects of rater behaviors are elicited by different methods.

In this study, I compared ratings of oral proficiency, accentedness, and comprehensibility awarded by English-as-a-second-language (ESL) teachers and American undergraduates to understand the role of rater background in measures of ITA speech. I also explored factors to which raters attended while they evaluated ITA oral proficiency.

The research questions guiding this study included:
1. Do ESL teachers and American undergraduate students differ in the severity with which they evaluate potential ITAs’ oral proficiency, accentedness, and comprehensibility, respectively, and if so, to what extent?
2. What factors draw raters’ attention while the raters evaluate potential ITAs’ oral proficiency? Are different factors more or less salient to different rater groups?

To answer the research questions, I had 13 ESL teachers and 32 U. S. undergraduate students evaluated 28 ITA candidates’ oral responses to the Speaking Proficiency English Assessment Kit (SPEAK), whose results are used to screen ITAs at Michigan State University. Raters assigned ratings online and provided written comments regarding the factors they took into consideration while judging candidates’ oral proficiency. After raters completed their ratings, one-on-one follow-up interviews were conducted to probe raters’ reasons for making
rating decisions. The scores assigned by the two groups of raters were subject to different statistical analyses to determine whether the two groups of raters differed in severity. Written comments and interview protocols were analyzed to understand raters’ rating orientations.

In response to the first research question, the results suggest that the rater groups did not differ in the severity they exercised when they evaluated the examinees’ oral proficiency. This finding is backed by the overall results of the multiple quantitative analyses, including the descriptive statistics of the raw scores, the classification of ITAs assignments, the FACETS analyses, and the Mann-Whitney $U$ tests. However, there were significant, between-group differences between the two groups’ ratings on accentedness and comprehensibility. The undergraduate raters were more severe when they judged the examinees’ foreign accents. They also perceived a significantly higher level of difficulty in comprehending the examinees’ speech.

The second research question delved into why raters with different backgrounds may differentially rate the speech of ITAs. By coding the written comments, I identified six main rating categories the raters reported they employed: linguistic resources, phonology, fluency, content, global assessment, and other, nonlinguistic factors. Raters’ attention to the first four rating categories was broken down further. For example, within the linguistic resources category, raters made comments on the examinees’ use of grammar, vocabulary, expression, and textualization. Within the phonology category, the examinees’ pronunciation, intonation, rhythm and stress, and foreign accent were all sources of attention. As far as fluency is concerned, raters judged the responses based on the repetitions or self-repair patterns and the speech rate of the speakers. In terms of content, raters noted whether the examinees fulfilled the task requirements, the ideas that the examinees produced, and the organization of the responses. Nonlinguistic factors included test-taking strategies, voice quality, and examinees’ emotions.

The quantitative comparisons of the written comments and the qualitative analysis of the interview protocols further helped determine the extent to which rater groups differed in the rating criteria they utilized. The results of these separate analyses converged, indicating that the ESL teachers and the undergraduates attended to several aspects of the linguistic dimensions in the examinees’ speech differently. Specifically, the results suggest that the teacher raters commented more frequently on a variety of linguistic features than did the undergraduates. The undergraduates, on the other hand, appeared to evaluate the examinees’ oral performances more impressionistically. The interview data reveal that many undergraduates were not familiar with the rating criteria for judging the SPEAK examinees and, thus, they made their rating decisions solely through their appraisal of whether they felt a particular examinee was qualified to be an ITA, or whether they would like the speaker to be their TA—a criterion not on the rating rubric. In either case, the data appear to suggest that undergraduate raters consider their personal feelings, perhaps even their fears, and their possible future experiences as students in ITA classes in judging ITA speech. They may tend to err on the side of caution and be more severe on accent and comprehensibility, regardless of oral proficiency, in anticipation of possibly having the test taker as a teacher in the future.

One pedagogical implication of this study has to do with the training of undergraduate students with regard to how to listen to accented speech. Many undergraduates have a general tendency to feel anxious about listening to foreign-accented speech due to their limited experience interacting with L2 speakers, or their lack of confidence in their own abilities to communicate or understand foreigners. However, it is not fair to say that ITAs are always at fault when there is a communication breakdown in the classroom. ITA programs should consider not only offering pronunciation instructions to ITAs but also make available training workshops that
teach undergraduates how to listen to and process accented L2 speech. These workshops can help reduce undergraduates’ anxiety while they listen to or converse with L2 speakers or their ITAs. Even through very limited training, undergraduate students can increase their ability to comprehend accented speech and enhance their willingness to talk with L2 speakers.

This study also has implications for ITA testing. Undergraduates are important stakeholders in the ITA testing context and should be included as a part of the ITA screening process. As the results show, although the ratings of oral proficiency assigned by the undergraduates were comparable to those assigned by the ESL teachers, significant differences on the ratings of accentedness and comprehensibility across the rater groups were found in this study. The undergraduate raters were more severe in terms of accent and comprehensibility judgments. It can be argued that the ESL teachers’ judgments of the examinees’ oral performances in terms of accent and comprehensibility were more lenient than the undergraduates’ because the ESL teachers paid more attention to specific, linguistic features in the speech samples, while the undergraduates tended to base their ratings more on accent (and for the undergraduates, the heavier the accent, the worse the comprehensibility of the speech) and overall feel. A main finding of this study is that undergraduates may not be able to act as impartial judges, even with extensive training, because they have something at stake—the possibility to be taught by ITAs who they cannot understand. This study’s results suggest that ITA programs should avoid having undergraduates as official raters, but rather use them to check the threshold of what undergraduates may consider as incomprehensible speech. On the other hand, ITA testing program should not underestimate undergraduates’ abilities to adapt and comprehend ITAs whose speech falls within that “grey” zone (between what undergraduate raters would call incomprehensible, but what expert ESL teachers would call comprehensible, since research has shown that through very limited training, undergraduate students can increase their ability to comprehend accented speech and willingness to talk with L2 speakers (e.g., Derwing & Munro, 2009; Derwing et al., 2002). Therefore, ITA testing programs should not fear such a gap. Nevertheless, the potential difference in severity between the official ITA testing raters and the undergraduates should still be constantly monitored, carefully evaluated, researched, and controlled.
References


Pickering, L. (2004). The structure and function of intonational paragraphs in native and


