LTJ_30_2_Rater_Bias

[Start of recorded material]
Moderator: From the University of Leicester in the United Kingdom, this is Glenn Fulcher with another issue of Language Testing Bites. In issue 32 of Language Testing we publish a paper by Paula Wink, Susan Gaff and Carol [unintelligible 00:00:21] on rate of second language background as a source of bias in the assessment of speaking. We have known about rata bias for a long time, and regularly take measures to control for it. But it is only recently that research into the source of bias found its way into the literature. In this podcast Laura Ballard stands in for me, to ask Paula Wink and Susan Gaff about their research into rata bias, as recording took place on site at Michigan State University where they all work. I would like to thank not only the authors but Laura and the Michigan technicians for their work on this podcast.
Interviewer: Thank you for coming on Language Testing Bites to discuss rata bias with readers of Language Testing.
Respondent 1: Thanks for having us, we appreciate it.
Interviewer: Language testers and educational measurement researchers have known about rata bias for a long time. Perhaps we can ask you to describe what tasks readers are asked to perform in speaking tests, and what rata bias is.
Respondent 1: That's a great question. Susan Gas, Carol [unintelligible 00:01:26] and I are very interested in this question. What we would like to look at is the ratings process and the outcomes of that, so we are interested in when ratings are changed, they learn how to orient themselves to the tasks that are being assessed, in a speaking assessment the characteristics of speech. When raters are trained then they rate the speech samples, and we hope that there is no bias, that they are not rating any characteristic that is not on the rating [unintelligible 00:01:58]. It is something like a bias could be raters who are harsher, giving worse scores to people who are females because they for some reason don't like a high pitched voice. That would be unwanted variation in test scores, related to some characteristic that is not on the [unintelligible 00:02:18].
Interviewer: You say that bias can be caused by a number of different factors. Can you outline these factors and say which you have found to be the most serious?
Respondent 1: Well researchers have looked at bias to see what factors. Some of the researchers have looked at occupation, the level of English teaching experience, how that affects their scores that they give. And also their familiarity with the accent. Familiarity can be in the accent of the speaker taking the test, and accent can be measured in different ways, so exposure to the test takers L1, or what we were interested in our study is even a low level of familiarity in terms of exposure to the accent through foreign language instruction.
Interviewer: One of the factors you mentioned is the linguistic background of the readers. In what ways can this affect how the reader awards the score to a candidate in a speaking test?
Respondent 2: Well this actually gets into our particular study. What we did in our study was look at how a language that a rater knows, how knowing a language either because it is spoken in their home when they were growing up, so-called heritage language, or because they studied the language as Paula just mentioned. Or even because they lived abroad. Knowing that language can influence a rater as they are evaluating speech samples. So we know that comprehension is facilitated when one is familiar with a particular foreign language, or even when a rater may know foreign languages in general, is used to hearing non-native speakers of their language, for example in the classroom. So that can facilitate comprehension. What we set out to do was to find out how that might translate into the situation where they have to rate speech samples such as the kinds of speech samples that are given in standardised tests, TEFEL being the one that we looked at. What we found was that there was such a bias when we matched them up with their native language of the test taker, and the particular experience that the rater had in a particular language, we matched that up. And we did find that there was more leniency for raters with Spanish experience, and with Chinese experience. They were more lenient when they were rating Spanish

native speakers, and Chinese native speakers. We also had a group of Koreans and the same was found in terms of there was greater leniency, although for Koreans it did not reach the level of statistical significance, although we do think that may be in part to the small sample size that we had for Korea.

Interviewer: Of great interest to many people listening to this discussion is the methodologies that we might use to investigate rata bias. I suppose the multi-faceted [unintelligible 00:05:30] analysis has been the most commonly used tool. Can you talk a little bit about the methodologies we have at our disposal and how they are used, and perhaps give one or two examples.

Respondent 2: It is true that [unintelligible 00:05:42] analysis with item response theory is one of the most popular methods because with that, as opposed to classical test theory, you can have a larger number of speech samples and a large number of raters. But the raters don't have to rate every single speech sample, so you can do a design where there are some common items that all raters rate, and use that to balance the scale across the different raters, and rating the different speech samples. A lot of the studies that have appeared in Language Testing have looked at rata bias in terms of [unintelligible 00:06:17] measurement. But there are other ways too, and we are starting to see more studies that are looking at qualitative data as well, so Susan Gaff and I have done a second study with this data where we invited the raters to come in and be interviewed through a stimulated recall protocol where they watched themselves rating. We asked them what they were thinking at that time. Sometimes they do discuss their rata bias in relation to accent familiarity. We have a paper coming out that uses that methodology, and takes on quarterly in 2013.

Respondent 1: It was interesting also to follow up on that with some of the comments that they made – oh that reminded me of my grandfather, something like that, so these would be heritage language speakers. When they hear this speech it does conjure up images for them, and it was through that particular methodology that you could find that out. You wouldn't know that if you were just looking at the [unintelligible 00:07:17].

Interviewer: I suppose the primary goal of this research is to understand rata bias so that we can either eradicate it or control it in some way. How do we use the findings from rata bias studies to make performance testing more fair?

Respondent 1: Well comes of it is just knowledge, finding out more about raters. What languages they have studied, how well they might know them, what might be heritage language, who might be a heritage language speaker. I think knowledge is important, and then bring that into rater training programmes and discuss it as part of what is not to be included in the [unintelligible 00:07:57]. It is really just gaining more information about the raters, and then openly discussing it and determining to what extent heavy accents, for example, should or should not be included in the decisions that we make.

Respondent 2: I think one of the theoretical questions is how can we tease apart for raters the problem of disassociating problems in pronunciation with problems in identifying and deciphering characteristics that come from accent. So those are very hard to distinguish but it is something that needs to be discussed in rater training programmes, because accent isn't supposed to be part of the rating [unintelligible 00:08:38] but sometimes it is when people have a hard time understanding because of the accent.

Respondent 1: It is hard to say that you can tell people not to conjure up images of their grandfathers. But then what they do with that information, not to translate it into a particular scoring issue.

Interviewer: Improving test fairness through reducing or eliminating rata bias is clearly a critical issue for testing speaking in particular. I'd like to thank you for taking the time to explain some of the issues at stake.

Respondent 1: Thanks for giving us this opportunity.

Respondent 2: We enjoyed talking to you.

Moderator: Thank you for listening to this issue of Language Testing Bites. Language Testing Bites is a production of the journal Language Testing from Sage publications. You can subscribe to Language Testing Bites through iTunes, or you can download future issues from ltj.sagepub.com or from languagetesting.info. So until next time we hope you enjoy the current issue of Language Testing.

[End of recorded material]