Language Testing Bytes **Issue 8**

Confidence scoring of speaking performance: How does fuzziness become exact?

An interview with Tan Jin & Barley Mak

From the University of Leicester in the United Kingdom, this is Glenn Fulcher with another issue of *Language Testing Bytes*. In issue 29-1, we publish an article by Tan Jin and Barley Mak on the application of Fuzzy Logic to scoring performances tests. Tan is a PhD student and Barley Mak, an Assistant Professor and Director of the Centre for Enhancing English Learning and Teaching, Department of Curriculum and Instruction in the Faculty of Education at the Chinese University of Hong Kong. In this article, they call their approach Confidence Scoring, and we are delighted when they agree to join us on *Language Testing Bytes* to explain a little bit more about their research and its potential application to our field.

Glenn: *Tan and Barley, thank you for agreeing to join us on Language Testing Bytes to talk about your article on confidence scoring in issue 29-1 of the journal.*

Tan: Thanks, Glenn. Thank you very much for inviting us to make this podcast.
Barley: Glenn, thank you. We would also like to extend our thanks to the anonymous reviewers for their insightful comments and suggestions. Thank you.

Glenn: *Well, thank you for that. Our reviewers will definitely hear that, but we will pass it on anyway. So the starting point of your article is a recognition that in assessing speaking performances, raters find it difficult to distinguish between adjacent levels on a scale, and that there is an overlap between different scales. Can you briefly explain each of these problems for our listeners?*

Tan: Sure. These two problems are actually related to the two elements of rating scales, namely, levels and scales. For example, we have level 1, level 2, level 3, and so forth. We have grammar scales, vocabulary scales, and sometimes content scales. In scoring a speaking performance, a rater first decides which level a candidate should be at. The rater then assigns levels in terms of different scales. These levels on different scales are usually averaged to produce a final score. However, it seems to be very difficult to actually locate a clear boundary between the adjacent levels–between level 1 and level 2, between level 2 and level 3, etc. Raters experience the same problem when they are asked to give only one exact score to a candidate's spoken performance. This is a point a number of studies have echoed in our field. It has been found to be very challenging for a rater to distinguish between adjacent levels, such as between level 4 and level 3 and between level 3 and level 2. This is a problem of indistinction between adjacent levels. Next, we come to the problem of an overlap in scales. When combining the scores on different scales to a final score -- an exact final score. Previous studies have also indicated that there is an overlap between scales. Let me give you an example. When we look at the content scale, we might judge how well a candidate is able to express their ideas with appropriate details. However, if a candidate cannot convey their ideas with sufficient detail in some way, it might also be because the candidate doesn't know the relevant words or how to pronounce and use the words. This overlap between scales shakes the foundations on which the averaging procedure in terms of calculating an exact final score are built. So we conceptualize the two problems as fuzziness in scoring speaking

performance. One is the indistinction between adjacent levels, while the other is the overlap between scales.

Glenn: *Right. You refer to the problems that you just described as "fuzziness", and in this paper you draw on fuzzy logic to model the problem. Before we talk about the approach you and your colleagues have taken can you first of all explain what "fuzzy logic" is, and why you think it is relevant to the problems you've identified?*

Tan: Okay. According to the Stanford Encyclopedia of Philosophy, fuzzy logic emerged as the proposal of fuzzy set theory by Professor Lotfi Zadeh in 1965. I'd first like to use a popular example here to briefly explain what fuzzy logic is. We may say a person is "tall" if their height is more than 180 cm; otherwise we'll say they are "short". We have defined the truth of "tall" as "more than 180 cm". In traditional dual logic, a statement is true or false, and an element either belongs to a set or it doesn't. So, from the example of height, a person who is 179.99 cm is not considered to be "tall". As in dual logic, we have no choice in determining whether person is tall or short. Ok, but in fuzzy logic, we give degrees of truth ranging between 0 and 1. So, a person who is 179.99 cm might be considered to be 0.9 "tall". This is more reasonable.

Let's return to scoring speaking performance. Raters are required to assign an exact score to a candidate's performance based on the rating scales. So we can say that the assigning activity is 'crisp' in nature because raters have no alternative but to say "yes" or "no" to a particular level (level 4, for example, as described in the rating scales). However, there is no clear boundary between adjacent levels as I just mentioned. So, fuzzy logic tries to model the problem as to what degree a candidate performance belongs to a particular level. So finally, we use membership functions and rule bases to solve the two problems of "indistinction" and "overlap". We've outlined membership functions and rule bases in our article, but I don't think I have the space to go into it here. You can refer to Zimmermann's 2001 book on fuzzy set theory and its applications if you are interested in the two concepts among other concepts.

Glenn: *Thanks for that explanation. So let's come to the real heart of the paper, which is the notion of a confidence score. Would you tell us what a confidence score is, and how it is arrived at by a rater?*

Tan: Sure. Confidence scores are awarded by raters based on the levels of certainty they have when they match candidate performances with descriptions of different levels on a scale. They may be absolutely sure or they may be not sure.

In scoring spoken performances, raters can use a score from 1 to 10 to show their scoring confidence in two adjacent levels. For example, they can give a score of 3 to level 4, and a score of 7 to level 3, meaning that they are more confident that level 3 is a correct level. But they feel that level 4 might be incorrect instead. Or they can award a score of 4 to level 3 and a score of 6 to level 2. The sum of the confidence scores should be 10.

If the rater is very confident about which level the candidate should be, the rater just assigns a score of 10 to that level. Raters then award their confidence scores on different scales respectively -- for example, the rater may first assign confidence scores on the Pronunciation scale, on the Grammar scale and then on the content scale.

Glenn: *As you just mentioned, confidence scores are awarded by raters based on the level of certainty they have when awarding a score. Perhaps you can say a little bit more about the notion of 'rater confidence' and how it features in your approach to scoring?*

Barley: Sure. Rater plays a very important role in scoring speaking performance, They actually have three roles. First, they must interpret the rating scales. Second, they must compare the description of each level with the actual candidate performance. And, third, they must make decisions themselves. Scores are thus awarded to candidates based on rater judgment of candidate performance.

As a teacher educator, I have been training teachers to score speaking performances. Although I provide detailed explanations about the rating scales as well as typical benchmark samples of each level, teachers – from my own experience, no matter how experienced they are – might still find it very challenging to distinguish candidate performance between adjacent levels. So I ask teachers participating in the scoring: what would you do if you have to make a decision and you find it difficult to distinguish between adjacent levels? The teachers' answers are generally that, first, they first hesitate between the adjacent levels, and second, they then choose the level which they feel a bit more confident with, although sometimes it's hard to say which level they actually feel most comfortable with and most importantly, most confident with.

So this is where our 'rater confidence' approach fits in. It serves as a possible solution to these constraints and the hesitation that raters find themselves in when being forced to choose between levels. Rater confidence is based on the level of certainty raters have when making judgements as to which level a candidate should be at. To make confidence scoring more user-friendly, confidence scores are assigned using a sum of 10 to adjacent levels, for example, 4 and 6, 3 and 7, and so on. In addition to this decision making aspect, raters do the same things they have in the past--they read the rating scales, and compare candidate performance with the rating scales. The only difference rests on the decision-making part. By adopting confidence scoring now, raters are allowed and encouraged to reflect their confidence in matching a candidate's performance to one particular level by assessing confidence scores! If the rater finds it very difficult to distinguish between level 3 and level 4, the rater can give a score of 5 to level 3 and a score of 5 to level 4 as well. So if the rater feels absolutely sure that the candidate should be at level 4, the rater can give a score of 10 to level 4. I believe putting confidence into the decision making is very important. And that's why we have named this approach "confidence scoring".

Glenn: *Thank you for that. I think that bring us to the question about what you think the advantages of confidence scoring are when using traditional rating scales to score speaking tests?*

Barley: Well, as Tan mentioned early on, a number of studies have highlighted two problems with traditional scoring practices. The first one relates to the indistinction between adjacent levels and the second relates to a kind of overlap between scales. Our confidence scoring approach acknowledges a rater's confidence in assigning scores of two adjacent levels to different scales. We see this approach as a possible solution to the problems we just mentioned.

Confidence scoring is actually based on and developed from traditional scoring. As I just mentioned, if raters are very confident in their judgments, they just score speaking

performance in a traditional manner by assigning a score of "10" to a particular level. You can see that confidence scoring embraces traditional scoring but it goes one step further. Because it provides a more flexible way of acknowledging raters' scoring confidence when making decisions. This is the first point I'd like to make.

Second, to deal with the overlap between scales, the confidence scores for different scales are processed into exact scores based on rule bases. So let me elaborate a bit on what rule bases are. Rule bases are established on the basis of specialist expertise. For example, if a candidate gets a level 3 for Pronunciation, and a level 2 for Grammar, the composite quality of the Pronunciation and Grammar scales will be determined on the basis of specialist expertise, rather than simply as an average of the two levels. The rule bases make the best of human expertise in dealing with the sophisticated composite quality of speaking performances.

Another point I'd like to add is that confidence scoring asks raters to assign more numbers to show their confidence rather than just give a single band score, as is the case with traditional scoring. For example, two candidates A and B are both given a level 3 under traditional scoring. Under confidence scoring, Candidate A might be given a score of 4 in level 4 and a score of 6 in level 3; Candidate B could be given a 1 in level 4 and a 9 in level 3. As you can see, the two candidates have the same band score using traditional scoring, but they have different scores under confidence scoring. It is preferred and possible to better discriminate between the two candidates using confidence scores. In this connection, we can see that confidence scoring may better discriminate candidates who would be at the same level when using traditional scoring.

Last but not least, as we put it in our article, confidence scoring is a new approach in scoring speaking performance. We hope further studies can be done to provide more evidence for enhancing the confidence scoring process we have been pioneering.

Glenn: *I'm sure that we are going to hear more of this in the future. Thank you very much for joining us on Language Testing Bytes to explain a little more about this fascinating area of research. I'm sure that it will be of great value to the readers of the journal as they tackle your contribution to issue 29-1.*

Barley: Thank you very much for giving us this chance to talk more about confidence scoring.
Tan: Thank you very much.

Glenn: Thank you for listening to this issue of *Language Testing Bytes*. *Language Testing Bytes* is a production of the Journal Language Testing from Sage Publications. You can subscribe to *Language Testing Bytes* 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.