**LTJ 27 3**

[Start of recorded material]

Interviewer: From the University of Leicester in the United Kingdom. This is Glenn Fulcher with another issue of Language Testing Bytes.

Issue 27(3) of *Language Testing* is a special issue on the automated scoring of speaking and writing tests, guest edited by Dr Xiaoming Xi. In this podcast we invited Dr Xi to talk about why automated scoring is such a hot topic, what the possibilities and limitations are, and of course, why the media and educators often get hot under the collar about computer algorithms scoring learner writing or speaking. On the technical side, for this podcast, I'm very grateful to Jeff Johnson who stood in to ask the questions on my behalf, and to Bob Scanneller [SL 00:57 for arranging local recording.

JJ: Could you tell me why is it that you think automated scoring is such a hot topic right now?

Respondent: Well, I think both the potential that automated scoring offers in terms of improving assessments, and also its controversial nature, have generated a lot of interest in the field. And automated scoring started about probably 50 years ago in the 1960s, and the purpose was to make it feasible and efficient to score a large number of essays in standardised tests. And we all know that over the past several decades of years [SL 01:34], several decades of automated scoring has made a lot of improvements to support assessment and learning. We all know that automated scoring can reduce the score turnaround time, significantly could reduce scoring costs to make tests more affordable to test takers, and when the automated scoring is used in conjunction with human scoring it has the potential to improve the reliability of the scores.

But, I think what really has put automated scoring in the spotlight is that in recent years there has been an increasing use of automated scoring in large scale language tests and some of these tests are very well known, so that has attracted a lot of attention. And, of course, you know, in some cases the use of automated scoring is very controversial, which has generated a lot of debate as well.

JJ: How does automated scoring of writing or speech work?

Respondent: Well, it's usually a multi-step process to try an automated scoring model. You start with a very large number of speaking or writing responses that have already been scored by human raters, and then these responses are analysed by computers for attributes or features, for example, grammatical accuracy or vocabulary diversity, all these different aspects of speech or writing that you're familiar with. And

then, you select and combine features in a principal way that can produce scores that are similar to human scores.

So, this multi-step process gives you a scoring model, a formula, which can then be applied to new responses, new speech samples or writing samples to get the scores.

JJ: If I understood you right, what you're saying is that automated scoring is developed based on human scores and automated scores are then compared to the human scores in order to demonstrate the validity of automated scoring. Did I understand you right?

Respondent: Yes, correct.

JJ: Okay, so then I have a question, is it a lot of testing organisations claim that the main rationale for using automated scoring is to overcome unreliability of human scoring. Isn't there a contradiction in logic there?

Respondent: Well, I these two statements do seem a little bit contradictory don't they, and there's actually truth in both statements, and let me explain why. The accuracy of automated scores is impacted by the quality of the human scores used for training the automated scoring model. So, everything else being equal if you have higher quality human scores your automated scores are going to be more accurate. While human scores can be inconsistent or unreliable for a variety of reasons, we all know that, but you can actually improve the quality of the human scores used for training the automated scoring model, by either, you know, conducting very rigorous rater training, or averaging the human scores on the training samples.

So, and also another point that I want to bring up is that, when you train an automated scoring model, automated scoring is actually based on the common patterns that emerge in numerous responses, in the scores of numerous responses, scored by a large number of raters. So, once you have a model and then the computer could apply it consistently to either verify the accuracy of individual rater scores in future scoring jobs, or you could use it to complement, use it as the second rater to complement human scoring. So, that's why I say, you know, automated scoring has the potential to improve the reliability of the scores.

But, I do want to emphasise that score reliability is only part of the story it's not everything. The consistency of automated scoring is not equal to, for example, validity or accuracy, don't confuse these concepts. An automated scoring model could score everything very consistently but using the wrong criteria or looking at only some of the things that human raters evaluate.

JJ: Whenever the media get hold of stories of automated scoring, it seems like one of the things they like to do when they hear that, oh, now, research scientists are trying to use computers to evaluate essays, or evaluate writing. One of the things they do, I've noticed, is they take speeches of great politicians, or they take novels from inspirational authors, or authors that are generally considered to be great, and they feed them into this computer scoring programme and then they report on how poor the scores in the feedback appear to be. And then, of course then, I guess the implication of that is, hey, look these things can't be any good because here's Hemingway, and look how this thing scored Hemingway. Do you think that's a fair comparison, do you think that's a fair test of it, and if not, you know, what mistake are they making, what's wrong with that argument?

Respondent: Well, that's a very natural response from the media, right, you know if I didn't have any knowledge of automated scoring I would probably do the same thing. The evaluation method and also the result really appealed to the lay public who doesn't know anything about the technologies behind it, but remember most of the current automated scoring systems for writing are actually developed to assess the writing of non-native speakers, or developing native speakers, not developing native writers.

And also, these systems are sort of designed to look at only a few limited types of writing, or genres of writing such as academic writing or expository writing. So, these systems are very sensitive to the kind of errors, problems and issues of developing writers, but not to the problems of more advanced writer whose problems may be, for example, a lack of coherence or inappropriate voice for the audience, that kind of problem.

And the other point, you know, the kind of features useful for assessing academic writing may not work for creative writing at all, so it's really not appropriate to use automated scoring, automated writing systems that are currently used to assess, for example, the language of inspirational speech writers, or inspirational novelists, because their language might contain sort of spoken features or dialogue features which are not standard written features.

So, I think to be fair you really have to look at how the system does in terms of the target population based on the right types of writing.

FF: My understanding is that automated scoring of speaking is a lot more complex than scoring of writing, why is that?

Respondent: Well, the first obvious thing is that to be able to score speech you have to recognise it first, right. You have to recognise and process the speech first, and that is actually the biggest hurdle in automated speech recognition, and think about natural, spontaneous speech, it's which is unpredictable. So, it does create a lot of challenges in terms

of automated scoring, and also speech may contain errors, speech disfluencies, fragments, and these just add to the challenge, so that's why we think, you know, automated speech scoring is a lot more difficult.

And another thing is that when you score speech you have to evaluate, for example, fluency, prosidy and pronunciation, which are particularly not considered for writing, so you have extra things to look at there.

JJ: What are some key research issues that the developers of these automated scoring systems, what are some issues that they face in developing them, in improving them, in making them useable?

Respondent: Well, I think there are a few things, the most urgent one, I think, is try to expand the kind of attributes or features that computers can analyse, for example, content, coherence. I know that computers can analyse vocabulary grammar and spelling with reasonable accuracy at this point, but, I think, to go forward we really need to expand the kinds of things that computers can do.

And the second thing is, automated scoring has a lot of impact on, you know, the way students prepared for the test and on how test takers speak or write in the test. So, if a computer rewards, for example, longer essays, longer sentences, or lower frequency words, and then test takers are going to be prompted to produce what's expected by computer, to try to get high scores. So, we really, you know, want to look at sort of the interaction between the test takers and on the automated scoring system.

And a third major issue, I think, is to look at the consequences of using automated scoring especially its impact on teaching and learning.

JJ: Do you think then that we are ready to use automated scoring to replace human scoring in a high stake setting, in a setting where the test results have some kind of high stakes to them?

Respondent: Well, it really depends, it depends on whether you are replacing human scoring completely, or you are just replacing one of the two human raters scoring each task, and I think currently we are in a position to use automated scoring to sort of complement human scoring for high stakes purposes. But I want to emphasise that we are not in a position, we're not ready yet to use automated scoring alone for speaking and writing in high stakes decisions.

JJ: Why is that?

Respondent: Well, I say this for a few reasons. First, we know that automated scoring systems now can produce scores that are comparable to

human scores, however, they can't mimic human raters, they can't score, for example, coherence, logic or content the same way human raters do. And second, the current state of the automated scoring system dictates that it's very vulnerable to cheating and also test scheming, and I'm going to explain why in a bit.

And, thirdly, I think for high stakes purposes, you know, for assessment used for high stakes purposes we really need to look at the consequences of using automated scoring, not just the accuracy, we need to look at the consequences. So, you might think, well, as long as I show that my automated scores are similar to what human raters would assign I'm fine, because at the end of the day it's the scores that matter.

However, they are, you know, we have to understand that automated scoring can impact a lot of things in the assessment process, and many of which may, in turn, change the meaning and also the accuracy of the automated scores. For example, there's this whole test coaching industry whose priority is to, you know, try to help their clients which, you know, who are test makers to gain the test to get high scores, and computers, because they are programmed to look at specific things, to look at limited types of features that are very amenable to technology, so they are a lot easier to fool than human raters.

And, eventually, I think, you know, test takers tricking of the system is going to sort of result in scores that do not reflect their actual language proficiency, and I think that issue is going to surface once the test coaching industry has caught up. We know that language tests used for high stakes purposes, as you said, are used for making important decisions about test takers lives, so, and also they have a profound impact on teaching and learning, so I think, you know, all the test publishers have an obligation to use automated scoring responsibly and appropriately.

So, I just want to reiterate, I don't think, you know, the current state of automated scoring justifies eliminating human scoring completely for high stakes purposes. Maybe someday, you know, when computers could read the content of an essay, could score sort of the coherence and the logic of human language in a meaningful way, we'll be more or less ready.

JJ: So, where do you expect us to be with automated scoring in ten years, what's the way forward, what are some of the next steps?

Respondent: Well, I think at this point a sort of blanket rejection of automated scoring isn't going to help us move forward, nor is an uncritical acceptance of the technology, and I think what we do need is a healthy dose of scepticism, and what we need is informed and targeted criticisms that can push providers of automated scoring to

innovate and to move in the direction that is beneficial to the users, and developers and researchers from diverse disciplines have to work together to improve the state of the ark of automated scoring.

I think in ten years' time we will have made important headway in terms of using computers to evaluate the content and coherence of human language. Is it an attainable dream that computers could score, could read the content of an essay some day? I think it is. We may not be there in ten years but I think we will get there and I think in ten years' time computers will be able to score and first recognise spontaneous speech with good accuracy, and also we'll see much greater use of automated scoring in high stakes decisions to complement human scoring.

JJ: Thank you very much for taking the time to talk about this today.

Respondent: Thank you.

Interviewer: Thank you for listening to this issue of Language Testing Bytes.

Language Testing Bytes is a production of the *Journal of 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*.

[End of recorded material]

NOTES: [SL 00:00] Sounds Like