The psycholinguistic approach to speaking assessment

An interview with Alistair Van Moere

Aaron Batty
abatty@sfc.keio.ac.jp
SRB General Editor

Alistair Van Moere is the vice president of product and test development at Pearson's Knowledge Technologies Group, and is responsible for the development, delivery, and validation of their automated language tests, including the Versant test (previously known as PhonePass). Versant is a telephone- or computer-mediated, computer-scored test of speaking proficiency, available in six languages, including English. Although computer scoring is widely used for multiple-choice test formats, Versant is revolutionary in its use of computer scoring of speaking proficiency. Van Moere and his colleagues have written extensively on the validation of the Versant test and its methods of automated scoring (e.g. Bernstein, Van Moere, & Cheng, 2010; Van Moere, 2010), and he has most recently raised eyebrows with his calls for a "psycholinguistic" approach to speaking assessment from the pages of Language Testing (2012).

Prior to his employment at Pearson, Alistair was instrumental in the development of the Kanda English Proficiency Test at Kanda University of International Studies in Chiba, and drew from his experience with its speaking test in his PhD work under Charles Alderson at Lancaster University—work which won him the Jacqueline Ross TOEFL Dissertation Award in 2010 (a portion of which can be read in Van Moere, 2006).

We were pleased that Alistair was willing to take some time out of his busy schedule to talk to us about the psycholinguistic approach to language assessment, the Versant tests, and the communicative/psycholinguistic divide.

In your recent article in *Language Testing*, you advocate a psycholinguistic approach. What exactly does this mean?

The theory underlying the psycholinguistic approach is that the speed and accuracy with which learners process language, even relatively simple language, tells us as much about their proficiency as more complex, communicative tasks can do. We don't actually need to elicit a variety of complex structures on a variety of topics while pretending to be in a variety of interactional situations in order to predict how well learners can handle themselves in a conversation.

A psycholinguistic approach refers to a particular way of (a) eliciting, and (b) scoring spoken responses. Elicitation uses constrained question-and-answer tasks, such as listen to a sentence and repeat it. Scoring involves three elements: accuracy, fluency and complexity. Let me give you an example. I'll say three sentence fragments and then ask you say them back in a correct, meaningful order: "to the cleaner / I gave my baby / to look after". Right. Native speakers will make sense of the meaning in a few milliseconds and then reel this out with barely a thought. When it comes to learners, though, some will say a fragment or two then trail off because they couldn't construct the meaning, some will say it correctly but slowly while they are processing it, some will say it

correctly but stammer over the articulation of certain words because their cognitive resources are so busy concentrating on meaning or word order. Every delay, stammer, stretched syllable, mispronunciation, missed word or missed word-ending, provides insight into the learner's psycholinguistic competence. And there's actually an interplay between accuracy, fluency and complexity. For example, learners tell us that when they focus on articulation their pronunciation or fluency scores go up, but they find it harder to focus on meaning and their accuracy scores go down, and vice versa. Naturally, as complexity or sentence familiarity get harder, both accuracy and fluency suffer.

The important thing here is that the tester, by means of controlled input and expected output, has a very standardized arena in which to judge the test-taker's speech for accuracy, fluency and complexity. This is in contrast to communicative tests, where test-takers have more control over the situation and may use various strategies to obscure or enhance the appearance of their linguistic range.

And how do you score psycholinguistic assessments?

Well the first thing to note is that psycholinguistic assessments do not require automated scoring. Sure, automated scoring makes the whole thing easier, but a teacher can administer these item-types in a five-minute test and rate responses for pronunciation, accuracy, and fluency, either on the spot or from recordings.

In our Versant tests, we score learner performances using our automated models. In the example of "I gave my baby to the cleaner," for example, we would conduct extensive field testing using that item. That is, we would gather the responses from numerous native and proficient speakers who performed on the sentence, and the responses from numerous learners of all proficiencies and many L1s. The native and proficient speakers provide us with the parameters of acceptability in the utterance, i.e. where linking or reduction occurs, where it's acceptable to pause inter-word, and so on. The learner responses tell us how difficult or complex the sentence is based on the number of word errors or other mistakes. As we have large item-banks we use partial credit Rasch modeling to determine item difficulty estimated from response data provided during field testing. The whole approach is empirical and quantifiable, and provides very granular scores. I tried to provide a readable explanation on this in my 2010 paper.

[Further information can be found in Pearson Education, Inc., 2011a. –Ed.]

What exactly do you mean by language assessments with "automated scoring"?

It refers to artificial intelligence systems that have been developed to assign scores like a human rater. We score both spoken and written responses and we use speech recognition, natural language processing and statistical modeling techniques. We do it by capturing test-taker performances from field testing, and having them marked by a pool of expert raters. Then, we analyze the data to establish which features of the performance are associated with high or low scores, and ultimately train an engine to assign scores like a human rater.

It's quicker and more reliable than human scoring, as long as the test is properly designed. I explain how it's all done in a series of short video clips here:

http://goo.gl/zZAjP

Could you give our readers a quick idea of how you scale these data? What model(s) do you use? How does the computer assign scores?

Wow, those are big questions and it's difficult to give a brief answer, but here goes. The spoken Versant tests last anything from 8 to 25 minutes and produce scores on a 20 – 80 scale for each of 4 subskills: Sentence Mastery, Vocabulary, Fluency, and Pronunciation. The overall score is a weighted combination of these subcores. For pronunciation and fluency, a panel of expert human raters applies rating criteria on a 0 – 6 scale to each field test response. Speech processors extract information such as the speed and delay in the response, the position and length of pauses, the segmental forms of words, the pronunciation of word segments. These features are used in models to predict the pronunciation and fluency of new, unseen responses. Sentence mastery and vocabulary are a combination of Rasch difficulty estimates from constrained items and predicted human judgments on criteria with a 0 – 6 scale. Rasch logits and criteria judgments are combined and put onto a normal distribution with 20 – 80 scale, according to global large-scale field testing. It's explained more fully in the Bernstein, Van Moere and Cheng (2010) article.

The Versant test uses computer speech recognition to assess learner speaking proficiency. According to the Pearson website, this technology was developed in-house. Allowing that speech recognition software has improved greatly over the years, most commercially-available packages are still pretty bad at it. What makes you (and Pearson) confident that you are getting accurate readings from your software, when, for example, the technology behind Apple's Siri, which has been in development for 40 years, has trouble understanding anyone who doesn't have a North American accent?

There several differences between Siri and the Pearson system. One advantage our speech recognizer has over other systems is that we've optimized it for learner speech. It means that we've developed it to recognize the pronunciations that learners have, so for example we can credit the learner for the content of their speech and debit them for mispronunciations, and the automated scoring does not confuse the two.

But the real difference between Siri and Pearson's system is that Siri doesn't know what you are trying to say, whereas the Pearson system usually does. Not only that, but during our test development we have presented the test items to a sample of learners and gathered their responses, so we know how test-takers respond to each item and what kind of words they use. Some item-types elicit highly constrained responses (e.g. read this passage; repeat this sentence). In this case, we are not conducting speech recognition so much as alignment of the test-taker's response with an existing set of response parameters. Other item types require the test-taker to construct a response, but they are bound by certain input or topics (e.g. describe this image; summarize this lecture). In this case the recognizer can anticipate certain frequently used words and word strings.

At Pearson our approach is to design tests that are compatible with the automated scoring technology. That means we maximize the things that the machine is good at and minimize what it isn't good at. It means carefully selecting our tasks and content. Frankly, it sometimes also means restricting our construct. But that's not necessarily a big limitation; every test is a compromise between length, reliability, task variety and construct.

We have other proprietary techniques embedded in the processing of data which ensure a good match between human evaluations of speech and the machine scores. But the proof is in the validation data. Machine scores correlate with human judgments at around 0.97 and 0.98, so the system is highly accurate.

What do you say to those who criticize your approach to speaking testing as too "form-focused"? Isn't the ability to communicate one's ideas more important than correct grammar or the ability to parrot back something they have heard? In real life, no one asks you to do anything remotely like the tasks you describe in your most recent *Language Testing* article (sentence repeat and sentence build). Why bother testing something that no one will ever be expected to do?

A big section of the article on the psycholinguistic approach is devoted to explaining why these tasks are more reflective of real-life communication than most people realize. That's why I review the literature on memory, retrieval, and chunking. Language production is highly memory-based, and when constructing sentences we frequently re-use chunks of language that we've previously encountered in speech because it's more cognitively resource-efficient.

It's also a misconception that sentence repetition is the same as "parroting"; I doubt that many informed academics believe that anymore. There is a large body of literature on this in the field of second language acquisition (where it is referred to as elicited imitation) and the task is generally acknowledged to involve comprehension of meaning and subsequent reproduction of meaning through form.

That said, it's true that the psycholinguistic approach is form-focused. When I said earlier that every mispronunciation or missed word ending provides information about language proficiency, well, that drives some communicative language testers absolutely loopy. For them, spoken performance testing is about conveying a meaningful message or demonstrating communication strategies, and not about dropping a suffix or measuring how long test-takers grope for vocabulary. And I don't disagree. But what we've found is that even minor errors of form can tell us a lot about the speaker's automaticity with the language. They are powerful indications of processing competence and can be excellent predictors of language proficiency and conversational ability, from beginner to advanced. Versant tests have been likened to other predictive tests in the field of medicine. For example, checking blood pressure takes only a minute and predicts very well whether the patient would get out of breath walking up a mountain. This saves you actually having to follow them up a mountain with your clipboard!

So you think there is something wrong with the communicative approach to testing?

The communicative approach is a work in progress, and we need to keep researching it. For me, the biggest limitation is the lack of empirical evidence underpinning our models of communicative competence. The field has defined various abilities or competences, but there is no way to prioritize them in terms of importance or measure exactly how they interact. For example, take organization. Now that seems to me to be an ability that is borrowed from L1. If you have good organizational skills in L1 then you can transfer it over to L2. But if you don't have it in L1, it's hard to see how you can ever have it in L2. Now you might say that organization is language specific – for example, in China when you write an essay you don't write your thesis statement in your introduction, but rather hold it back until the conclusion, and sort of hit your reader with it as the culmination of your argument. But that is just a different form of organization that you can apply in L1 if you wanted to. It's not something that you could only apply in one language and not another language. So, much of organization is a language-independent skill.

Pragmatics is another area that I struggle with. It seems to me to consist mostly of grammar and vocabulary plus emotional intelligence. You are only able to express yourself in an appropriate register if you have the emotional intelligence to understand what is appropriate and the grammar to carry it out. So it appears to me that a large part of pragmatics consists of context-specific

grammar structures (form) in the target language plus a non-linguistic trait that is transferable across languages, in addition to a sociolinguistic ability of selecting the right grammar and vocabulary for the right situation. So, I think research needs to focus on identifying the contribution of the different components of CLA [(Communicative Language Ability) –Ed.] and finding out how much of language proficiency is attributable to core language skills (e.g. grammar, vocabulary, automaticity), how much is attributable to non-linguistic, transferable skills (e.g. critical thinking, organization, emotional intelligence) and how much is genuinely an ability that exists in the L2 illocutionary and sociolinguistic competences, irrespective of the L1.

How have your views on spoken language testing evolved?

Although I still believe in communicative frameworks to help us define constructs and design tests, my research experiences have led me to focus more on concrete, core skills that are essential for meaningful communication.

When I was at Kanda University we ran group discussion tests for thousands of students for placement and progress-monitoring. I thought it was a very useful communicative test—three or four students discussed a prompt and were evaluated by two raters using an analytic scale. The performances provided seemingly rich, social interactions. I researched this for my PhD, and as I conducted more analyses I gradually became doubtful about the whole approach. In a study that I co-authored with Miyoko Kobayashi, we found that a large portion of the score variance was attributable simply to the amount that students spoke, even after controlling for proficiency. So if two students were the same proficiency level, the one that spoke more scored significantly higher. In a study I conducted with William Bonk, we discovered that outgoing students scored higher than average or shy students. Gary Ockey subsequently ran this study more rigorously than Bonk and I, and he published it in Language Testing, but his findings were basically the same. Then during my PhD I looked more carefully at the score reliability and found that students' scores really varied quite a lot over consecutive test occasions, and so were too unreliable for anything but low-stakes decision-making. I also found through discourse analysis that the task itself elicited an alarmingly limited range of functional interactions.

So, I was getting despondent about these findings and having a kind of crisis of faith about our ability to reliably measure talk-in-interaction. At that time I started working on the Versant tests. I wasn't at all convinced at first. Like many people, my reaction was: "This is just repeating sentences. It's not assessing communicative skills." But it took me about a year to change my mind. I had to understand the scoring, and analyze the data for myself. And I saw the tremendous reliability, and also how the Versant scores were correlated with interview tests. Then I began to realize that we can complement the communicative approach with more reliable, more standardized tests. So that's why I advocate a psycholinguistic approach in addition to a communicative approach.

[For a full report on the validation and reliability of Versant tests, see Pearson Education, Inc., 2011b –Ed.]

Thank your for taking the time to speak with us. Your ideas are always intriguing. To wrap up, what can we expect to see from you in the future?

We are developing a four-skills test for young learners on the iPad which involves very interactive, tactile item types, including speaking and writing activities. We've just launched a 4-skills, computer-based placement test for universities which assesses CEFR A1 through B2, and which provides immediate scores. We are also in the process of launching a business English test for Ja-

pan and Korea, which is going to be just 90 minutes for all 4-skills, and will be practical, reliable, and very accessible.

In R&D we are also working on test security features such as speaker verification. This helps you ensure that the person sitting in front of you holding a test certificate is indeed the same person that took that test and earned that score. You could have them jump on the phone or computer, offer a speech sample, and we'd compare it to the test and took and confirm whether it's an exact match or not. There is a lot happening and over the next few years I've no doubt that we're going to wow people with automated scoring technology even more than now.

References

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355–377. doi:10.1177/0265532210364404

Pearson Education, Inc. (2011a). Research. *Pearson – Versant spoken language tests, patented speech processing technology, and custom text services*. Corporate Website. Retrieved October 15, 2012, from http://www.versanttest.com/technology/research.jsp

Pearson Education, Inc. (2011b). *Versant™ English Test: Test description and validation summary*. Palo Alto, CA: Pearson. Retrieved from http://www.versanttest.com/technology/VersantEnglishTestValidation.pdf

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, *23*(4), 411–440. doi:10.1191/0265532206lt336oa

Van Moere, A. (2010). Automated spoken language testing: Test construction and scoring model development. In L. Araújo (Ed.), *Computer-based Assessment (CBA) of foreign language speaking skills, Joint Research Centre scientific and technical reports* (pp. 84–99). Brussels: Publications Office of the European Union. Retrieved from http://publications.jrc.ec.europa.eu/repository/handle/111111111/15037

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325–344. doi:10.1177/0265532211424478