Language Testing Bytes Issue 22

Diagnostic Language Assessment

An interview with Eunice Jang

GF: Issue 32(3) of Language Testing this year is a special issue on the subject of diagnostic language testing. For as long as I can remember, language testing textbooks refer to diagnostic tests as one of the "five types" of language test. They then proceed to ignore it almost completely. But in recent years there has been a real surge in interest and research in diagnostic testing. One researcher who has focused on diagnostic testing in Dr. Eunice Jang from the University of Toronto, and she is a contributor to the special issue. We're delighted that she's agreed to talk to us about this really important field of study.

Welcome to Language Testing Bytes, and thank you for agreeing to share something of your expertise with our listeners.

EJ: Thanks Glenn for the opportunity. What a pleasure to talk with you!

GF: Just to make things very clear for our listeners, can we start out by asking you to provide us with a definition of diagnostic language assessment?

EJ: I must confess I am not a big fan of definitions! But let me take a stab at it. I see that as an assessment approach used to provide detailed information about the areas of strength and further improvement in user-specified language domains. While such diagnostic information can serve both teachers and students, its primacy is on learners who are the change agents of their own language learning. Misdiagnosis resulting in mistargeted treatments and agony vs correct diagnosis without actionable plans can be equally problematic, and both precision and actions for improvement are core features that define high quality diagnostic assessment.

GF: Now that's clear for everyone listening to us, let's start to explore the topic in a little bit of detail. In the introduction I mentioned the traditional "five types" of language test. The two that are routinely relegated to the side-lines are aptitude testing and diagnostic assessment. Can you tell us why there has been such a huge interest in diagnostic assessment in recent years? What's driving all this new thinking and research?

EJ: This is a great question. Although clinical and psychological diagnostic measurement has a long standing in the modern measurement history, diagnostic language assessment is relatively a recent development. I can think of three main driving forces underlying its popularity:

First of all, it has drawn much attention for the past 15 years, and this period coincides with neoliberal educational reforms based on economic models of accountability. A prime example of such is the No Child Left Behind Act of 2001, which was a testing-driven educational reform for school improvement based on yearly adequate progress all students make. Recent political interest prompted a new form of testing programs, that assess and track student growths and intervene learning more effectively. Developing a high quality diagnostic assessment is a serious endeavour in terms of its requirement for financial and professional resources. Whether it accidentally coincides with such reform movements, well, it's uncertain nor entirely desirable. But it appears that political climates provided a test bed for fertilizing diagnostic assessment.

Another critical contributor is significant shifts toward interest in assessing skills that enable the application of knowledge to tasks beyond assessment of 'knowledge'. This shift from assessment of how much students know toward what students can do with that knowledge reflects the societal, public, and policy interest in preparing learners to be competent in the 21st century. However, not all diagnostic assessments may focus on skills over knowledge, and I remain keen on ongoing dialogues about the matter.

Lastly, I think the biggest contributor is a conceptual shift toward pragmatism, that is, a recognition that assessment is a means to educational goal (which is essentially learning). This renewed interest calls for assessment that allows systematic documentation of evidence of growths in target domains. And it is well reflected in effects-driven assessment design by Drs. Fred Davidson and Glenn Fulcher and use-oriented test validation by Dr. Bachman.

I think all these factors, political, conceptual and pedagogical shifts, contributed to creating a sort of tipping point for diagnostic assessment, as Malcolm Gladwell may say.

GF: So, in your own research your make it clear that you're interested in the use of test scores – or perhaps more specifically, the outcomes of assessments, for the improvement of learning. In the United Kingdom, in particular, this isn't something new. The concept of Assessment for Learning (AFL) has been around for quite a long time, and the work of researchers like Black and Wiliam have become standard texts for trainee teachers. In what way does your approach to diagnostic assessment – as outlined in your contribution to the special issue - differ from what we traditionally think of as AFL?

[Here, I'm really thinking of whether your approach is radically different, whether it provides an alternative approach, or builds upon, what has gone before. This is really the only place where you can refer directly to your article rather than discuss the topic more generally, but don't get too technical though! We want the podcast to be accessible to a wider listenership! Also, remember length! The entire podcast cannot exceed 20 mins approx.!]

EJ: There is no doubt that AFL along with AAL (Assessment as learning) has changed the landscape of today's assessment and pedagogical practice. In particular, AFL provided a framework of reference for teachers and educators in communicating their daily practice with other stakeholders. Through AFL, I think teachers experienced conceptual empowerment and professionalization to some extent. AFL indeed opened quote unquote the "black box" of education, which is what goes into education, instead of focusing on what comes out of education.

However, AFL is rather a conceptual framework that articulates the role of assessment in its broadest sense. Diagnostic assessment definitely is an AFL, but not all assessments used for AFL are necessarily diagnostic assessments. There is a diagnostic assessment tool, but we don't say it's an AFL tool. So diagnostic assessment is a specific type of AFL that features a set of principles in terms of theories of specific skill development, task characteristics, scoring and communication.

In the paper published in the special issue, we focus on the last component, which is diagnostic feedback, the end result of diagnostic assessment, in the form of communication and actions. Information that is shared, negotiated and used by learners should be specific enough (in terms of target skills and domains being assessed), age and ability appropriate in terms of grain size, and should be based on cognitively engineered, diagnostically sensitive language tasks. We learned that how such diagnostic feedback is perceived, processed internally, used for future learning depends on a constellation of various cognitive, psychological and affective characteristics of learners as well as family and classroom environments they are in. This reminds us of the complexity of human mind, a need to take into account dynamic interplays among cognitive and non-cognitive traits.

GF: I'd like to ask a very specific question at this stage, if you don't mind. In the recent literature I've noticed that there is a tendency to try to retrofit existing proficiency tests with diagnostic information. I suspect that this is what Fusion theory was all about, and you've written about this in *Language Testing* before. Where do you now stand on this issue? Do you think we can get diagnostic information from traditional proficiency tests, or do we have to design diagnostic tests with the diagnostic purpose in mind?

EJ: Let me give listeners a quick summary of what you refer to in your question. My dissertation examined the validity of diagnostic assessment practice based on the retrofitted application of psychometric diagnostic model called Fusion Model (more recently RUM) to LanguEdge (a prototype test of TOEFL designed to provide support for teachers and students). While the study featured some positive outcomes of the given CDM and responses from users, I concluded with some warnings that tests designed based on traditional CTT or IRT principles may not provide diagnostically reliable information due to a lack of diagnostic discrimination at upper and lower ability ranges and a lack of balanced item distribution across target skills.

Where do I stand now? Hmm, honestly I must say I have had a pragmatic turn in that retro-fitted application of diagnosis modelling to existing testing programs is inevitable to some extent. But I think of it as a transition to the next era of assessment. And we have gained much knowledge and experience from the application.

Since 2005, I have had opportunities to retrofit CDM to K-12 achievement tests and more importantly develop a new diagnostic assessment using CDM principles. Like architecture (as you eloquently elaborated in one of your many papers, Glenn), a test's designed with an intended specific use. Proficiency tests whose primary goal is to discriminate among test takers for selection are built upon the psychometric principles of normal distribution whereas K-12 standards based achievement tests tend to serve as a minimum competency measure in which a score distribution tends to be negatively skewed. When CDM's retrofitted to these two different assessments, precision of diagnosis is compromised differentially. Actually the new diagnostic assessment based on CDM principles shows the highest level of diagnostic discrimination. Again, without retrofitting, we couldn't have learned what we now know about cognitive diagnostic assessment, and with that knowledge and skills, I believe we can now design operational diagnostic assessments.

GF: And what do you think is the future for computer based diagnostic assessment? Are researchers looking at automated diagnostics, or is it humans who are required to interpret the outcomes of computer mediated tests?

EJ: Oh, this is the question I get excited the most these days. Again, I approach the matter pragmatically, what will maximize use. Immediacy or timeliness of diagnostic feedback is critical for maximizing its use. To this end, computerizing diagnostic assessment is integral. I think research-based design of diagnostic assessment and reliable diagnostic scoring modelling based on a representative sample will allow for automatized diagnostic profiling and feedback delivery. There is a growing interest in machine learning through the pattern-recognition algorithms to unsupervised assessment data, but I don't think it will replace human endeavors, as diagnostic assessment of language learning and cognition should be grounded in mature theories and interpretations. I envision more active roles that humans can play in supporting the use of diagnostic information for furthering learning with an aid of computerized diagnostic assessment

GF: It's always difficult to cover a topic as broad as this in one podcast. So we could bring the podcast to an end with a little bit of speculation. Can you pick one or two ideas that you think we'll come back to again and again, and will characterize the nature and scope of diagnostic assessment research in the next five years? I know that's a difficult question, but perhaps you'd like to give it your best shot.

EJ: It's indeed difficult, so instead of speculative prediction, I'd like to answer that question by what I think shall further advance diagnostic assessment. I think the use of diagnostic assessment for learning will and should be a recurring topic. Both teachers and learners need to experience positive changes as a result of diagnostic assessment whether it's internally or externally developed. We need more empirical evidence accounting for the mechanism of change resulting from the use of diagnostic assessment.

In addition, I'd like to see a methodological diversification of diagnostic assessment. CDM is psychometrically powerful yet presents many challenges due to its rigid psychometric assumptions to be met and as a result is limiting in small scale local contexts of learning. I welcome research into psychometrically less constrained diagnostic modelling, such as latent class or profile analysis, clustering methods, or subscoring approaches. And we saw some exciting research using such profiling approaches for diagnosis at this year's LTRC conference by Drs. Ari Huhta and Charles Alderson.

Lastly, I think there will be research on diagnosing the development of non-cognitive traits that interact with target language competencies. I feel traditional definitions and about what's construct relevant or irrelevant will require reconceptualization along with ever-evolving validity frameworks. As the interactionalist view of language assessment posits, diagnosing language ability should take into account various contextual and noncognitive variables instead of controlling for their influence.

GF: Well, Eunice, I'd like to thank you for joining us on Language Testing Bytes, and for staring into your crystal ball for us at the end. I'm sure that your paper in issue 32(3), and the others in the special issue, will help define the research agenda for years to come. As we mentioned at the beginning of the podcast, you have specialised in diagnostic assessment for many years, and I'm sure that our listeners will rush off to download your article.

Listeners may also be interested in Eunice Jang's recently published book entitled Focus on Assessment Focus On Assessment: Research-led guide helping teachers understand, design, implement, and evaluate language assessment, in the Oxford Key Concepts for the Language Classroom series

EJ: Thanks so much for the opportunity, Glenn. I hope my two cents provided food for thoughts, not definitive answers. Thanks again.

GF: Thank you for listening to this, the final episode of *Language Testing Bytes*. We produced the first issue in 2010 to accompany Issue 27(2), and we were delighted that Mike Kane joined us to talk about Validity. The podcast has run for 5 years, in which we've gone on to produce 22 episodes. I hope you'll agree that all our contributors have been excellent, and added great value to the journal. I am rotating off the editorship of the journal at the end of 2015, and the financial situation being what it is, there are insufficient resources from the publisher to maintain this production. But all the podcasts will be maintained in an archive at ltj.sagepub.com, and on languagetesting.info. I'm sure this won’t be the end of language testing podcasting. So do keep your eyes open for future developments.