LTJ 28.3 Standards Based Testing

[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 28(3) of *Language Testing* is a special issue on the subject of standards based language testing in North America. This issue has been guest edited by Craig Deville, who is Director of Psychometric Services at Measurement Incorporated, and Micheline Chalhoub-Deville, who is Professor at the Department of Educational Research Methodology at the University of North Carolina in Greensboro.

Thank you to both of you for coming on Language Testing Bytes to talk about the special issue of the Journal that you've just guest edited on the subject of Standards Based Testing.

CD: Well, Micheline and I would initially like to thank you and Language Testing for giving us the opportunity to put together this special issue. In addition, we'd also like to thank the authors and the reviewers who contributed so much to this endeavour. Because the topic of the issue may be viewed by some as specific only to the US context, you, Glenn, went out on a limb in believing that the topics raised in this issue will be of value to the international readership of *Language Testing*, and we believe this as well, that although the NCLB legislation is particular to the US, many of the language testing policies and practices that have emerged [UI 01:30] will be of interest to the international language testing community. For example, Bailey and Long go into great depth discussing the construct of academic English and its relationship to content standard.

In another paper, Dorry Kenyon and his co-authors demonstrate how to build a vertical scale of English language development, which again has implications for how we view the language construct. In addition, Charlie Stansfield's paper details the implementation of accommodations on subject matter assessment, such as maths or science for English learners, and his work contains important repercussions from validation of test scores. So, as the readers of the Journal will see, all of the papers were written within the context of no child left behind, do contain information that will be applicable in a broader sense.

Interviewer: Thanks for that introduction. I often feel that there's actually some confusion over just what it is, so perhaps I could begin by asking you to explain just what standards based testing is?

CD: For sure. With respect to standards based testing we should keep in mind that the system requires both content standards and performance standards. So, content standards are intended to delineate what students should learn, know, be able to do, content standards serve, if you will, sort of subject matter blueprints for curriculum instruction and assessment. At the present time, each state within the United States has its own set of content standards, but in next year's most will be adopting the so called common core standards, which in essence means we will have national standards for the first time in the United States. Micheline and I in our introductory paper to the issue provide information as to what the Obama administration is currently doing in order to promote an agenda of, in essence, national standards and national testing.

So, I just mentioned content standard, performance standards on the other hand, specify how much of the content students should know or be able to do in order to be classified as, say, proficient. Based on performance test scores,

when I say performance I mean test scores, students are classified into proficiency categories, for example, into below proficient, proficient, or advanced. The accountability element comes in when authorities evaluate percentages of students who fall into the performance or proficiency categories, and then determine whether student performance meets stated expectations or criteria. Schools and teachers may be subjected to punitive consequences if not enough of their students attain prescribed achievement expectations.

By the way, sometimes content and performance standards are wrapped up into one scale, one such popular scale would be, here in the United States, would be the [UI 04:27] guidelines.

Interviewer: Yes, thanks. As you've hinted in your explanation, standards based testing seems to have grown out of an accountability agenda, and in the United States it appears to be directly linked, as you mentioned, to the no child left behind legislation. Can you just explain that link a little bit more for us and perhaps say just what role standards based testing plays in accountability?

MD: You're right, Glenn, the accountability agenda and the push toward standards based assessment in the US, are dictated largely by no child left behind legislation, NCLB. Basically, NCLB is the most recent reauthorisation of the Elementary and Secondary Education Act, ESEA, and that was an act back in 1965 [SL 05:17] under President Lyndon Johnson. And, this special issue, Mike Bunch does an excellent job laying out the historical development that led to the most recent reauthorisation of the Elementary and Secondary Education Act. So, I will not dwell on the background but I would like to say that ESEA was, and is, intended to govern primarily secondary education while permitting state and district to largely maintain control over their [UI 05:49].The Elementary and Secondary Education Act, which is physically [SL 05:56] reauthorised every five years basically promotes high standard and accountability.

The Federal government through NCLB, provides considerable money with the goal of achieving specific educational goals, and the government wants states and ultimately schools, of course, to be held accountable for how resources are spent. The push for accountability and the use of test scores as the primary measure for accountability have grown immensely, especially over the past 10 years, and we expect them to continue to grow, and as Craig mentioned, under President Obama, there have been some issues with the label, raise to the top, and we have about 4 billion dollars, over 4 billion dollars, being spent to push for further reform in education. Plans are underway for national standards and children related consortia based testing programmes are being developed as well.

[UI 07:01] sanction, I would say that up until recently schools faced sanctions if their student scores were not adequate, but now there is also pressure to use student scores to evaluate teachers, including decisions of hiring and firing, promotion and pay. Some states are even utilising what is called, value added models, which are statistical models that purport to ascertain how much a given teacher has added to the student education, as measured by test scores over time. And, there's a great deal of controversy about not only the quality of these models but how they are being used as well.

Interviewer: Yes, and as you said, there is a lot of controversy, and there's also a great deal of disagreement among language testers as to the desirability of these systems and the way in which they're being used. Can you tell us a little bit about just

where the battle lines are drawn, and what your take is on the usefulness of standards based testing as it's evolved in the US?

MD I guess we can say that the controversy emanates from the pros and cons of accountability based testing. Some of the objections include holding teachers in schools responsible for learning without consideration for the larger, societal factors that also impact learning. For example, poverty, home stability, family emphasis on education, school attendance for migratory families, and other such factors. Other objections include the narrowing of the curriculum where the focus is on what is being targeted [UI 08:48]. We're also seeing state [SL 08:52] sacrificing bilingual education to promote English language proficiency, and this is particularly damaging when our global reality emphasise the importance of being multi-lingual.

On the positive side, I would say that despite the many controversies, many language testers in the US would likely agree that one big positive consequence has been that English language learners now receive more attention than they ever did in the past. Before NCLB this group of test takers was too often excluded from subject matter assessment altogether, and there was no requirement to track their progress in learning English either. Now, English language learners, and their schools, are desegregated to that a clean picture can be obtained as to how this special group of test takers is performing relative to others, and relative to itself, over time.

The expectation is that this special group of learners, and indeed all the learners, are expected to make what is termed AYP, Adequate Yearly Progress. So, everybody is expected to be taken into consideration and attended [SL 10:12] to by the schools, including ELLs, and that's new.

So, people will disagree about how these [UI 10:21] are now used to make decisions about English language learners, but at least, the student population is no longer being neglected.

Interviewer: Okay, thanks for that explanation. What I'd like to do now is to move on to a couple of technical issues, and the first one is vertical scaling, which seems to be a central concern. Do you think you could briefly say something about what vertical scaling is, and how language testers go about doing this?

CD: Sure, I'd be glad to. To begin though, I want to differentiate between vertical scaling and horizontal equating. Most of us are familiar with horizontal equating, which is intended to take scores from different test forms and put them on a common scale. When you have multiple test forms, for example, forms A, B and C, which we would see with large testing programmes when tests are administered at different times and places, these forms are designed to be very, very comparable in terms of their measurement and giving construct domain [SL 11:26]]. Yet, they may differ somewhat in terms of difficulty. So, a given test form might be harder or easier than the other forms and so a comparison of scores, just raw scores, across these forms would advantage or disadvantage some examinees. So, equating, or horizontal equating, procedures are used to place scores on a common scale.

On the other hand, vertical scaling, notice that I'm not using the word vertical equating, and that's because I'm not talking about equating in a traditional sense, so vertical scaling focuses on tests or performances of students at different grade levels, where both the subject matter and the difficulty of the content change. The challenge with vertical scaling is how to address the differences in content or in curriculum across non-consecutive grades or levels.

Vertical scaling presumes that we can place learners across grade levels on a continuum and indicate their progression as they learn and develop their knowledge and

skills. We're all familiar with vertical scales actually, ACTFL and CEFR can be considered descriptive vertical continuum. And, as soon as we assign scores to these and link them, then in a sense you have a vertical scale. An underlying assumption of vertical scales is that the construct is a new dimension, meaning that we're measuring the same thing, let's say reading ability, when testing beginning or advanced language learners, when testing very young children or older teenagers.

In a language testing field we have relative consensus that the language construct is not uniform in terms of the components and their relative contributions to performance for different proficiency levels, and for different age groups. So, we need to be cognisant of the fact that while the students are put on one scale, that is we have a common metric for all, there are subsequent shifts in the construct make up at different levels. The psychometric assumptions that need to be met in order to construct the unified scale do not acknowledge these shifts. Therefore, it's not surprising that both content experts and psychometricians disagree amongst themselves as to whether such scaling is compatible with our view of language.

We need, perhaps, to think about having multiple overlapping scales that accommodate these shifts and allow us to make somewhat of a common metric system. Therefore, we really would recommend to readers to take a close look at the article by Kenyon and his colleagues. Dorry and his colleague group provide an excellent description of the many steps one needs to undertake in a word to construct a vertical scale, and, perhaps the most important step, one that is often disregarded or given short shrift, is the design of the test administration itself, and readers will see just how complex this test administration has to be in order to collect appropriate data for vertical scaling.

Interviewer: The other technical issue I'd like to explore is content. There appear to be some very complex approaches to mapping, test content and standards, and mapping both to the curriculum. This seems to me at least to be a huge change in the United States, in the past it was often argued that tests should be independent of any particular programme of instruction or text book, but now it seems that the test is almost completely integrated with programme content, so that teaching to the test is the same as delivering the curriculum. Is it possible for any single test to adequately sample an entire curriculum, and is this really a good thing for teachers?

MD: You ask some very good questions here, Glenn, and I agree with you that in the US the preferred [UI 15:32] to testing practices has been that proficiency based as opposed to achievement based testing. And, Craig and I have discussed such differences in terms of the Cambridge versus ETS [SL 15:46] practices in the chapter we wrote for the Educational Measurement publication in 2006. I would say that historically in the US, and in the interests of promoting meritocracy, testing programmes have shied away from [UI 16:05], because the argument was that such practice tended to favour elite schools and give unfair advantage to those who attended those schools. So, proficiency based tests, which are independent of any specific curriculum or instructional approach, as you said, have become entrenched here. So, the testing practices in US schools which meant [SL 16:30] a departure from that proficiency model, and I believe the change is, in principle, a positive one.

As language testers we usually are concerned about how our testing efforts complement the broader educational objectives of the learning and teaching contexts. We want more than just a test score to attach to a test paper, we

want the assessment process to enhance learning and promote good practices by educators.

A good standards based system will give good alignment among the components of the curriculum, the instructions, the assessments, so that teaching to the test entails teaching to the content standard, and if they are broadly defined that's a good thing. And here, I'd like to say that teaching to the test should never include teaching to the particular test, or items on a test.

I would like to add a cautionary remark here. With high stakes and high pressure accountability the purpose of teaching to the test mutates from a focus on learning to a focus on increasing test scores. The accountability pressures we're seeing has forced teachers and other educators to engage in questionable behaviour, and primarily at increasing test scores, and teaching to the test may be the least effective of such behaviour.

With regard to content mapping, you mentioned, I'd like to direct our listeners to the work by Bailey, who also has an article being issues. Bailey and her colleagues, Butler and Sato [SL 18:18], from UCLA, have provided the field with a very good example of how to develop systematic standards to standard linkages and they are typically used for the curriculum and for test design. And these targets, both language and [UI 18:35] standards and I think this is valuable work and I look forward to seeing more of it in the future.

Interviewer: Well, thank you very much. I'd like to thank both of you for guest editing a very important issue for the Journal, on a topic that's clearly going to be with us for a long time to come. This collection of papers is obviously going to become essential reading for researchers in the field. And I'd like to thank you both also for taking the time to talk to us on Language Testing Bytes, and helping readers to get to grips with what is happening in the United States.

CD: Well, again, we'd like to thank you, Glenn, for accepting a special issue of *Language Testing* comprising of policy and research investigation of the assessment of English learners under NCLB in the United States. We very much enjoyed working with you and the contributors to this issue, and we hope the issue is well received by the Journal's readers. Thanks again.

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

[UI 00:00] Unintelligible