Shiken: JALT Testing & Evaluation SIG Newsletter. 12 (2) April 2008 (p. 21 - 25)

Insights in Language Testing: An Interview with Carsten Roever

by Yi-Ching Pan
National Pingtung Institute of Commerce, Taiwan
The University of Melbourne, Australia

Carsten Roever is a senior lecturer in Linguistics and Applied Linguistics, School of Languages & Linguistics, at the University of Melbourne, Australia. He earned his PhD in Second Language Acquisition from the University of Hawai'i at Manoa in 2001. He has worked in test validation at the head office of Educational Testing Service in Princeton, NJ, USA from 2001 to 2002. In 2005 he published *Testing ESL pragmatics* and in 2006 he co-authored *Language testing: The social dimension* with T im McNamara. His research fields include second language acquisition, interlanguage pragmatics, language testing, and cross-cultural communication. This interview was conducted in person in March-April 2008.

What prompted you to get involved in the language testing field?

My undergraduate background was in TESOL and psychology, and language testing and second language acquisition seemed like the best way of combining the psychometric part of psychology with language. Also, during my undergraduate and master's programs at the University of Duisburg in Germany, I worked a lot with Christine Klein-Braley, who invented the C-test, and through her I got into this field.

Can you describe how your thinking about construct validity and face validity has changed since first entering the profession?

To me, construct validity is absolutely the most important consideration in any kind of test development. Since I really got into understanding Messick during my PhD, that's what I came to believe. Of course, the social dimension plays a role as well, particularly in terms of test consequences. I see a bit more of a role for face validity now than I used to. In the past, I tended to think that face validity was completely irrelevant. There is a certain relevance in that if people don't believe in the testing that's been done with them, they might be resistant to it and ultimately perform below their competence. For the psychometric testing conditions to be fulfilled, participants must perform at their maximum ability. Still, face validity is a minor consideration compared to construct validity. However, for people who are not in the testing profession, it represents how they look at the test. Because of this, we need to make the test look appealing, so it gets accepted, but under no circumstances can we ever compromise construct validity to enhance face validity.

Why are current validity theories unable to adequately account for the social dimension of language testing?

Well, it's not that the theories themselves are unable to account for it. Messick and later researchers like Kane, for example, discuss this issue at length. The real problem is that no one has suggested practical procedures for investigating it.

A lot of procedures to have been suggested investigate the internal aspects of test validity, most likely because that is an area with which test validation researchers feel comfortable. Psychometricians like statistical procedures. You can't investigate the social dimension simply through statistics. Such an investigation requires other procedures, and that reason could account for the minimal amount of work conducted so far, because there are very few suggestions on how to conduct such research. The establishment of procedures to validate the social dimension of language testing is a very complicated issue. Many more qualitative and ethnographic approaches are required. To get a picture of how tests impact society will most likely require (often costly) longitudinal studies. Few of us have enough patience to observe and monitor such a phenomenon for, say, periods such as ten years, but I'm sure some people will. Many of the social effects of testing become more obvious longitudinally. This type of research involves a more historically-oriented process. In addition, what applies to one context may not apply to others, as qualitative research recognizes context-dependency. Often the research methods can be used across various contexts. However, since the results cannot easily generalize and transfer between contexts, the findings can be very, very different depending on a constellation of contextual factors.

"The establishment of procedures to validate the social dimension of language testing is a very complicated issue."

On page 75 of the text that you co-authored with Tim McNamara in 2006, you raised a question about cross-cultural performances. At this point in time, do you feel that there is a construct such as the "aptitude for cross-cultural performance"? If so, what do you believe would be the best way to measure that construct?

Yes, there is, but it's a tricky construct. You can at least hypothesize there is such a thing as a larger or lesser ability to successfully engage in cross-cultural communication. People who are successful at that, I would think, would tend to have certain characteristics, and so they might, for example, be more sensitive to miscommunication. They might feel that something is going wrong and try to repair it. They might have a stronger understanding of the effect of the context on language use. They may acculturate more easily and therefore be able to at least participate in the roles in which other cultures would place them. The line here would be somewhere between linguistically-oriented factors and personality-oriented factors. A lot of research in the area of "cross-cultural adaptability", as some people call it, is about personality factors and there is a question whether people who are ethnocentric could do this. That's one thing. There are also linguistic pragmatic factors attached to that. There has been almost no investigation about them, but I think there are possibilities here.

One controversy in language testing seems about the extent that Rasch modeling can or cannot validate a test. Some people seem to approach validity from a technical perspective in terms of "infit". Others (such as Lassche, 2007) question whether depending on Rasch theory for larger validation issues might lead to a sort of tunnel-vision. What is your stance on this and what do you see as the strengths and limitations of Rasch modeling?

"Rasch analysis in and of itself does not buy you validity. It only shows you whether an item fits or doesn't fit, but just because the item fits, that does not mean that the item measures the construct."

Rasch statistics allow you to see to what extent an item or person fits the expectations of the model, and that is what is referred to as "item fit" or "person fit". Rasch analysis is very valuable as an item analysis tool to supplement -- or even supersede -- Classical Test Theory item analysis. But Rasch analysis in and of itself does not buy you validity. It only shows you whether an item fits or doesn't fit, but just because the item fits, that does not mean that the item measures the construct. You need other approaches for that. Messick outlined them in substantial detail, as have other people. Just doing Rasch analysis does not show you whether an item measures what you want it to measure.

In terms of strengths and weaknesses of Rasch theory, I should say that I'm not a "Raschi", and I haven't studied Rasch analysis in great depth. I've studied Item Response Theory, but primarily with people who were two or three parameter adherents, so my own tendency is more towards two parameter IRT. Generally, I don't think that the guessing parameter in 3 parameter IRT makes a substantial difference. Rasch analysis has an important advantage in that it can very easily handle ratings, so if you are dealing with ratings like in speaking tests, you have to do a bit of acrobatics to fit that with normal IRT, but with Rasch there is no problem at all. However, Rasch doesn't look at discrimination, and discrimination does make a difference. In reality, you are limited by your sample size anyway, and one parameter IRT or Rasch modeling needs much smaller samples than two parameter IRT. You can do Rasch with one hundred people. For two parameter IRT, depending on how the sample behaves, you probably need three hundred, five hundred, or more. Given the small numbers in most of our field, with the exception of large testing organizations, Rasch is probably one of the more feasible approaches and quite a powerful analytic tool. I'm not saying Rasch analysis is good or bad. It's useful but it's certainly not a cure-all, and it certainly does not tell you whether a test or an item is valid or not.

When you teach graduate courses on testing, what points do you feel are often most challenging for students to grasp? What materials have you found especially helpful in teaching graduate-level testing courses?

There are different ways of teaching graduate courses on testing. You can teach them very statistically and focus on the psychometric side of it, or you can teach them much more conceptually and focus on the methods of testing various skills, looking at it more from a construct validity perspective. There are challenges to overcome in both areas. If you go the statistical route, then the challenge is statistics, of course. Unless students already have a background in statistics, they should probably take at least one introductory statistics course. If you go the more conceptual route, the theories behind validity and construct validity are not easy bedtime reading. The writing of Kane and Messick is very dense and difficult. For students who have very little experience in testing, it can be very, very theoretical, and very much dissolved from reality. It would probably help quite a bit if either they already have experience or are involved in some test development effort as they are doing the class so they have

a way of applying what they learn in class. However, that's not as important if you take a statistical approach, because then you can just analyze existing data sets.

What books are helpful depends on your general approach. I always like Hughes's *Language Testing for Language Teachers* because I think it is a good introduction. J. D. Brown's *Testing for Language Programs* is also quite good. There's another good one published in 1995, *Language Test Construction and Evaluation* by Alderson, Clapham, and Wall. Tim McNamara's *Language Testing* represents a more conceptual approach. Of course, it's quite abbreviated because it's the shortest introductory book, but that can be a good starting point for broader readings.

I think for people to really get into language testing, I would probably teach a two-semester course - first introducing language testing, and then exploring advanced issues in language testing - because to digest so many things in one semester would be quite a challenge. If many graduate students are also language teachers, it would also be a good approach to combine testing and teaching practices. I think that could be, for example, a project component of a testing class. They could be asked to think about the context that they themselves work in and some of the tests that they either use or should use, and then look at those from a theoretical perspective or develop a new test given the material that has been covered in class. That's probably much more useful in the conceptually-oriented than in the statistically-oriented class.

What projects do you have on the horizon?

I can think of four projects, and there are others floating around that tend to pop up. I'm looking at testing pragmatics in an ESL placement test with our Language Testing Research Centre (LTRC). We integrated an implicature part in our placement test and we want to see how that relates to the rest of the test. Measurement of pragmatic aptitude is something I'm actually currently working on with the Center for the Advanced Study of Language (CASL) at the University of Maryland. We first defined the construct and are now developing the tools and hope that these those two aspects will sharpen each other. Something slightly different from the previous two projects is work on discourse completion tests. That reflects my pragmatics interest side, particularly research methods in interlanguage pragmatics. I'm also very interested in pragmatic development in less commonly taught languages and I'm currently working on putting together and editing a collection on the pragmatics of Vietnamese as a native or target language.

Works Cited

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation.* (Cambridge Language Teaching Library). Cambridge & New York: Cambridge University Press.

Brown, J. D. (1996). Testing for Language Programs. Upper Saddle River, NJ: Prentice-Hall.

Hughes, A. *(2002). Language Testing for Language Teachers.* (2nd ed.) Cambridge & New York: Cambridge University Press.

Lassche, G. (2007). Rasch & quality control: Controlling data, forgetting quality? In T. Newfields, I. Gledall, P. Wanner, & M. Kawate-Mierzejewska. (Eds.) *Second Language Acquisition - Theory and Pedagogy: Proceedings of the 6th Annual JALT Pan-SIG Conference.* May. 12 - 13, 2007. Sendai, Japan:

Tohoku Bunka Gakuen University. (pp. 42 - 55) Retrieved March 28, 2008 from http://jalt.org/pansig/2007/HTML/Lassche.htm.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University

McNamara, T. & Roever, C. (2006) *Language testing: The social dimension*. (Language Learning Monograph Series) Malden, MA & Oxford, UK: Blackwell Publishing.

Roever, C. (2005). *Testing ESL pragmatics*. Frankfurt: Peter Lang.