LTJ 27 2

[Start of recorded material]

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

In this first Language Testing Bytes podcast we're very lucky to welcome Professor Mike Kane. As many of you will know, Professor Kane recently moved to Educational Testing Service to take up the Messick chair in test validity. He's a prolific author on this topic and wrote the influential chapter on validation in the fourth edition of Educational Measurement. He's taught at the University of Illinois at Urbana-Champaign, the State University of New York at Stonybrook, and the University of Wisconsin, Madison.

Prior to moving to ETS he was Director of Test Development at the National League for Nursing, Vice Principal for Research and Development at ACT, and Research Director at the National Council of Bar Examiners.

Mike, welcome to Language Testing Bytes. It's a pleasure to have you talk to us on this very first issue of the podcast.

Respondent: Well, I'm glad to be with you.

Interviewer: Right, your approach to validation is widely quoted in the *Language Testing* literature, and in combination with evidence centre design it formed the basis for the revision of the test of English as a foreign language. Now, in my reading of your work it seems to me that the notion of the interpretive argument is central to your approach. Can you tell us what an interpretive argument is, and perhaps give us an example?

Respondent: Okay. I think many of my views, like everybody else, come from my experiences and I've spent a lot of my working life in different applied settings. I worked in… my first job was in a placement in proficiency testing programme at the University of Illinois, and subsequently I was Director of Test Development at the National League for Nursing, and then I worked for ACT on a variety of things. So, I've worked in a lot of applied settings. It seemed to me that the way to [UI 02:21] in looking at validity more theoretically as the appropriateness of interpretations and uses, that one of the differences between these different settings is kind of the interpretations we use than the kinds of uses that we were putting the test to.

So, it seemed important to me that we should clear about what the proposed interpretation and use were, and what the implications of those were, before we launched into analysing the validity of these interpretations and uses. So, I put forward the interpretive argument as a way of specifying the proposed interpretation and use in some detail and explicitly, and I don't know that an argument is the only way to do it, I think in some contexts construct language helps, and in some cases content specifications are really important. But, to sort of make it general it seemed to me that the notion of argument was a good one.

So, in one area in which this came up was in [UI 03:32] verification testing. There's a tendency for people to want to show that the test predict performance, and that really doesn't work very well because when somebody gets licenced as a doctor or a nurse they go off and they work in lots of different settings, yet the people who fail the test don't get to practice and the criteria for evaluating how well somebody's doing are kind of fake [SL 03:57]. So, it's really hard to do predictive validity but, on the other hand, it seems that [UI 04:03] verification serves a really useful purpose in establishing that somebody is well enough qualified to be allowed out into practice at a specific point in time.

So, one of the places where this came up for me was that in the interpretation and use of [UI 04:21] exams you don't have to include a notion of predicting performance in the future, the argument works quite well if you say, that we're going to evaluate somebody's competence in a domain of knowledge, skills and judgement at a particular point in time. And if they meet certain standards we'll give them this piece of paper that they can hang on their wall and they can practice, and if they don't meet those standards then we consider them to be not well enough prepared to be let loose on the public, and prediction never comes into it.

So, in some cases I thought that we were making the job of validation too difficult by not specifying clearly what the intended interpretation was and [UI 05:14]. And, in some other cases, we weren't being hard enough because there were all sorts of inferences being made on the basis of the test scores that weren't addressed in the validation. So, my intent was to sort of… and my basic concern was to find a general method for specifying the proposed interpretation and use to put it on the table for evaluation. So, that was the general idea.

Interviewer: In your explanation there it seems pretty clear when you were referring to the argument that you kind of draw on the work of Toulmin, and I've always been struck by the similarity between Toulmin's notion of substantial argument and Dewey's concept of warranted assertion. Both seem to require that we provide reasons of warrants and evidence to support the claims we make, but also postulate that any conclusions we reach are contingent upon future research that challenges the interpretive argument. Do you see your work as drawing on these insights, or is there an evolution in this pragmatic approach to inferencing?

Respondent: Well, I think, historically, it doesn't because I wasn't really very familiar with that work. I've read a little bit of Dewey and I always found him very insightful but he didn't really have a big impact on forming my idea. As an undergraduate, early in my graduate career, I was a Physics students so I came more out of the philosophy of science, people like [UI 06:57] and particularly Lackitose [SL 07:00], and my advisor in graduate school was Patrick Soupisse [SL 07:04] who is a logician and philosopher of science.

So, I came that way but I also recognise that sort of the strict logical mathematical kinds of models that people use in these contexts in many cases, and that are seen as being particularly desirable didn't really work for testing programmes in applied settings, that they were much messier and context dependent. And, one of the things I liked about Toulmin is that he sort of, his treatment of argumentation is very general and flexible, and he allows for the fact that the criteria that you use for warranted assertions might vary from one context to another, that is the kinds of evidence that you would expect to be able to get and use and one context might be different from that in another.

I do think though I'd like to learn more about Dewey and the pragmatists but I haven't done that in the past, so.

Interviewer: Well, moving on from that and the interpretive argument, I mean some of our listeners may be more familiar with the term validity argument, which is used a lot in language testing, than interpretive argument, and in your approach to validation you use both, so could you explain briefly what the relationship is between an interpretive argument and a validity argument?

Respondent: Okay, I see the interpretive argument as sort of stating the proposed interpretation and use of sort of putting it on the table for discussion evaluation criticism. I then see the validity argument as producing an evaluation of that interpretive argument of whether the test and the context and the inferences being made actually fit together in a way that makes sense, is defensible, and, to use that favourite word, or warranted by the available evidence.

I sort of see that the… I think of the interpretive argument as being the one little thing that I add to this enterprise in terms of terminology in that I was looking for some way of describing this system for specifying what's being claimed so that it could then be evaluated in the validity argument. And I think the validity argument notion I sort of got from Krombach 1988 paper, and from Howse's work on programme evaluation.

I think by the way that, as an interesting aside, that Krombach views on validity fit in with his notions of programme evaluation, because programme evaluations are intrinsically messy and context dependent, you know, its action programme is out there in the world and is trying to achieve something, it had some consequences. And, there are a lot of assumptions being made and outcomes that can be hard to measure and quantify, so it's not a neat mathematical, logical system which we tend to like in measurement theory, sort of mathematical models. And I think Krombach and Howse, sort of that, those aspects of programme evaluation come into their notions of validity and I've sort of bought into that to a large extent. I like the [UI 11:02] of their work.

Interviewer: Well, I'd just like to follow-up on that and your definitions by asking about the place of constructs in your approach, which follows on rather nicely there. I recently read a 2010 paper by Carol Chapelle and colleagues in *Educational Measurement Issues and Practice*, entitled *Does an Argument Based Approach to Validity make a Difference*, and they say that they found the approach that you've just described to be a very practical and pragmatic approach to structuring a test design project, one in which they couldn't clearly define a construct and operationalise that construct. Now, what they say and I'm going to quote this to you is, "Kane's organising concept of an interpretive argument, which does not rely on a construct, proved to be useful." Do you think that we can bypass construct definition in the way that they imply is possible?

Respondent: I think we can. One of the problems, I'm not really crazy about construct language mainly because construct now in our literature has so many different meanings that unless you specify what you mean by a construct at the beginning of an article nobody would know what you were talking about if you discussed constructs. In the original version in Krombach and Meale [SL 12:29] a construct is a theoretical entity embedded in a sort of mathematical or at least a formal theory and the construct is implicitly defined by the theory, something like, you know, entropy in thermodynamics, or something else.

Now, we don't have constructs like that. We don't have theories like that in lots of areas in which we work, and a few we have something that approximate it. And that notion sort of was associated with a logical positiveness to some extent, and is less probably central to philosophy of science then. So, then constructs became, sort of, general ideas of what we're talking about and were defined in terms of any connection with anything, get these statements in literature where any connection between your variable and any other variable is grist for the mill in validity, and it makes it really difficult to know what to do, you know, if there are infinite number of possible things you could look at and no criteria for saying which one you should do first, you tend to grind to a halt.

So, I think construct language helps in some areas. I think if you're talking about lots of psychological attributes, personality characteristics, like aggressiveness in someone, construct language makes sense and you can do something that, sort of, approximates Krombach and Meale to some extent. And that was the area in which Krombach and Meale introduced the concept to deal with. I mean they explicitly said that they introduced construct validity to deal with measures of attributes, like aggressiveness or intuitive ability or something, that couldn't be defined in terms of content domains and couldn't be defined in terms of predicting some criteria.

So, in some areas I think it works, in some areas I think defining content domains makes sense, and all of those can be subsumed under the general notion of an interpretive argument. So, I prefer not to focus so much on construct language, mainly because I think it has so many meanings that it's meaningless then.

Interviewer: Yes, I was going to say that from there it falls on nicely to the issue of content, which has recently come back into the fray as it were. So, if I can just turn to this recent controversy, I mean, in 2007, Lisette and Samuelson published this paper, *An Educational Researcher,* in which they argued that the whole of validity should be about content, which is, I guess, a move away from construct language as well, and, together with reliability they, kind of, argue that it should be, at least should be, classified as concerns that are internal to the test. As I understand it in this conception everything else that we would think of as Messick's aspects of validity would essentially be excluded from validity.

Now, I know that you've responded to this in a number of papers already but could you tell us why you find the Lisette's definition of validity to be problematic.

Respondent: Okay, I think on purely logical grounds I would say it's not necessarily problematic. You can define validity any way you want, right, I mean we can find a term, we can stipulate that a term means X, Y or Z. So, in a particular context if you wanted to define it in terms of content based evidence and reliability that would be okay, and I've talked about what I call observable attributes that are sort of defined that way, as one [SL 16:39] category.

But I think there are lots of things that we measure that go far beyond that, they make assumptions about underlying trends [SL 16:48] that make assumptions about what's going to happen with relationships that other variables predictions of future performance have consequences, etc. So, the meanings that we assign to our test scores often go far beyond content, and I think if we restrict validity to just content what tends to happen is the sort of version of begging the question, you know, we define validity in terms of a very restricted definition content and reliability, and then we make all these other inferences and we say, well, we have a valid test so those inferences are going to be okay, implicitly, and I think that doesn't work.

Basically, I think that the other thing that Samuelson and Lisette say is that, well, these things like consequences and fairness, and so on, will be dealt with elsewhere and then my question is, well, where? You know, who's going to do it if we don't do it? The local user doesn't have the resources to address these issues. Politicians typically are not interested in addressing what they don't want to address them because they might find out something that would be inconvenient. So, if we in measurement don't address them, who, is going to do it, and when? And my sense is that, never.

So, pushing, the other thing, I guess one other thing I'd say is that I think validity has evolved to be broader than the definition by Lisette and Samuelson, and if we want to go back to that we probably need to rewrite a lot of our text books and other things, because we have statements saying, validity is the most important characteristic of testing programmes, and, if it's only content then I would say, in some cases, it's not the most important context, the most important issue. You know, in some issues, again, in placement tasks, the consequences whether you put the person in the right course is really the central issue, it's not the content of the test per se.

Interviewer: Well, on that note, I can see your contribution to Issue 27(2) of *Language Testing*, which was a response to a paper on the topic of how we can define and achieve fairness in testing, and I guess this would be fairly, it would be, you know, well outside of the definition of validity that Lisette's is arguing for. He would definitely put uses and the consequences outside of validity. So, what place do you think fairness and consequences have within your approach to validity?

Respondent: Well, I tend to think that consequences have really been a part of our notion of validity for a long time, especially in placement selection tests etc., so the consequences in them is thoroughly new. What is a bit new is social consequences which got pushed, again I think that was implicit way back but Messick made it explicit and pushed it very hard in a number of articles and sort of brought it to the foreground.

But, I think that we should take responsibility for the consequences of our testing programmes, and one of the basic rules, as I understand it, for physicians in their oath is that they should do no harm, and I think in testing it's hard to be able to guarantee that you're going to do no harm, because if you assign somebody to a particular programme sometimes you're going to make the wrong decision, and if you hire one person and then not another there's always a trade-off there and somebody has lost. But, I think that we have an obligation like everybody else in every area to think about the consequences of what we are doing and try not to have larger negative consequences as well as having, sort of, immediate positive consequences of what we do.

So, I think consequences need to be thought about, and we should think about them, whether you want to include that in validity or as something else is a debateable question, but it seems to me a natural place to put it or at least to have it, it could be in other places as well, but it's a natural place to have it because consequences for individuals and for employers in terms of, or educational institutions in terms of selection and placement have been an integral part of validity for a hundred years. And so, adding the larger social consequences seem to me not to be a tremendously big leap.

Now, the difficulty with dealing with consequences is that it pulls us out of the, there are a lot of trade-offs and social, and even political, concerns in dealing with consequences, so it's a messy, sort of, difficult question, and it pulls us away from nice mathematical models, but I don't think that's a bad thing.

Interviewer: And, as you say, it's a debate that is going to go on for some time, not only in the pages [SL 22:37] of language testing but conferences and hopefully beyond and this, I hope, will contribute to that. But, I'm afraid, Mike, on that note it's all we have time for in this podcast. So, I'd like to thank you for this interesting conversation and for giving all the readers of *Language Testing* such a great insight into your approach to validation.

Respondent: Well, it's been a pleasure talking to you. Thank you.

Interviewer: Okay, bye.

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