Keynote presentations to the International Association for Educational Assessment (IAEA) 2008 Annual Conference

Sylvia Green Research Division

The 34th IAEA annual conference, hosted by Cambridge Assessment, took place in Robinson College, University of Cambridge from September 7th to 12th. The main conference theme was *Re-interpreting Assessment: Society, Measurement and Meaning*. The conference was the largest IAEA conference ever with around 500 delegates from 58 countries; 130 papers and 8 posters were presented. The highlights of the event were the two keynote presentations by Professor Robert Mislevy and Professor Dylan Wiliam.

PROFESSOR ROBERT MISLEVY:

Some implications of expertise research for educational assessment

The first keynote was presented by Professor Robert Mislevy, Professor of Measurement and Statistics at the University of Maryland. He was previously Distinguished Research Scientist at ETS and is a member of the National Academy of Education. He has been president of the Psychometric Society, and received career awards from the National Council on Measurement in Education and the American Educational Research Association. His research applies developments in technology and cognitive psychology to practical problems in assessment. In his address Mislevy focussed on the implications of expertise research for educational assessment commenting that developments in psychology and technology had led to exciting times in assessment. He provided insights from his research and their implications for assessment design. His descriptions of complex cognitive processes were presented in a way that was accessible to his audience and he provided meaningful illustrations of the concepts he introduced. He outlined difficulties in the contexts of cognitive processing limitations and knowledge, describing expertise as 'the circumvention of human processing limitations' (Salthouse, 1991). He also explained Walter Kintsch's theory of reading comprehension where relevant patterns from long-term memory may be activated in some contexts and not in others. In this theory the writer depends on many patterns and conventions that have developed over hundreds of years through the interactions of billions of people in the form of the letters, the syntax, and the words of the language itself. Mislevy explained that the physical and social context, what we have been working on, our purpose for reading, even the time of day, can influence our comprehension. He went on to draw on Kintsch's theory in relation to expertise research.

In expertise research into cognitive task analysis experts and novices are compared in replicable conditions and questions are asked, such as, *What knowledge is needed? How is it represented? How is it used? What makes tasks hard?* He suggested that experts organise their knowledge effectively and that they perceive, understand and act in terms of fundamental principles rather than surface features (Chi, Feltovich and Glaser, 1981). He emphasised the importance of interaction with a situation and of external knowledge representation for information processing and cognition. A key question posed was – *How do you use improved understanding of the nature and acquisition of expertise to design and conduct assessments?* Other important questions were raised, for example, *What complex of knowledge, skills and other attributes should be assessed? What behaviours or performances should reveal those constructs? What tasks or situations should elicit those behaviours?* (Messick, 1994). He drew on a socio-cognitive perspective, looking at four important aspects of expertise:

- Organisation of knowledge
- Knowledge representations
- The importance of interaction
- Social aspects of expertise.

He went on to describe examples of computer assisted assessments with a range of design tasks and simulations and considered implications for task design and design patterns. The differences between experts and novices were set out in three contexts: using disparate sources of information; formulating problems and hypotheses; vocabulary and language usage. He concluded that insights from expertise research can improve the practice of assessment and support deeper learning and that doing so requires a deeper understanding of assessment design. He also suggested that suitable conceptual frameworks, tools and exemplars are now beginning to appear.

He concluded that assessment is – structuring situations that elicit evidence about students' thinking and acting in terms of patterns 'in our head' and that many of these patterns are social in their construction, acquisition and use. Some of the insights derived from cognitive psychology and developments in technology can be applied directly in familiar forms of testing while others need to be developed outside traditional and familiar practices, in technological environments. Mislevy's final comment was a challenge to assessment professionals: 'We, as the community with interests in learning and assessment, are moving to our next level of expertise.' This was a thought-provoking presentation that was well received by delegates. It set out some interesting thoughts that were referenced in different contexts throughout the conference.

PROFESSOR DYLAN WILIAM:

What do you know when you know the test results? The meanings of educational assessments

The second keynote address was presented on the final morning of the conference by Professor Dylan Wiliam, Deputy Director of the Institute of Education, London. In a varied career, he has taught in urban public schools, directed a large-scale testing programme, served a number of roles in university administration and pursued a research programme focussed on supporting teachers to develop their use of assessment in support of learning. He posed an intriguing question – 'What do you know when you know the test results?' In his presentation he explored the meanings of educational assessments and focussed on the conference theme, *Re-interpreting Assessment: Society, Measurement and Meaning*. He addressed a number of fundamental issues:

- The importance of (un)reliability
- Evolving conceptions of validity
- (Mis)uses of assessments for educational accountability
- Some prospects for the future.

He discussed the difficulties of using classical measures of reliability and of ensuring the accuracy of scores as well as the precision of grades. He proposed that a test is valid to the extent that it assesses what it purports to assess and that the key properties in terms of content validity are relevance and the extent to which the test is representative. He also discussed issues of content validity, criterion-related validity (concurrent and predictive) and construct validity and he identified three important issues related to validity:

- Validity is a property of inferences, not of assessments.
- The phrase 'a valid test' is therefore a category error ('like a happy rock').
- Reliability is a pre-requisite for validity.

He proposed that validity subsumes all aspects of assessment quality including reliability, content coverage, relevance and predictiveness, but not impact. He identified threats to validity in terms of inadequate reliability as – construct irrelevant variance and construct under-representation.

Wiliam then moved on to some practical applications. The first of these was in the area of school effectiveness. He asked, 'Do differences in student achievement outcomes support inferences about school quality?' In discussing this he referred to the threats to validity that he had previously identified. This led to a further exploration of the threat of construct irrelevant variance and construct under-representation. In outlining the social consequences of inadequate assessments Wiliam referred to the Macnamara Fallacy:

> *The first step is to measure whatever can be easily measured.*
> *(This is OK as far as it goes).*
>
> *The second step is to disregard that which can't be easily measured or to give it an arbitrary quantitative value.*
> *(This is artificial and misleading).*
>
> *The third step is to presume that what can't be easily measured really isn't important.*
> *(This is blindness).*
>
> *The fourth step is to say what can't be easily measured really doesn't exist.*
> *(This is suicide)*
>
> (Handy, 1994 p.219)

He also quoted Goodhart's Law – *All performance indicators lose their meaning when adopted as policy targets* – and he gave examples of this including inflation and money supply as well as national and provincial school achievement targets. He warned against the effects of narrow assessment because of the tendency to create incentives to teach to the test and to focus on:

- Some subjects at the expense of others
- Some aspects of a subject at the expense of others
- Some students at the expense of others.

He concluded that reliability requires random sampling from the domain of interest and that increasing reliability requires increasing the size of the sample. He suggested that using teacher assessment in certification is attractive as it would increase reliability in relation to test time as well as increasing validity by addressing aspects of construct under-representation. However, he identified problems of a lack of trust ('Fox guarding the henhouse'), problems of biased inferences resulting from construct-irrelevant variance, and the potential introduction of new kinds of construct under-representation.

In his concluding remarks he outlined 'the challenge' – to design an assessment system that is:

- Distributed – so that evidence collection is not undertaken entirely at the end
- Synoptic – so that learning has to accumulate
- Extensive – so that all important aspects are covered (breadth and depth)
- Manageable – so that costs are proportionate to benefits
- Trusted so that stakeholders have faith in the outcomes.

He extended this challenge to delegates whilst recognising the size and complexity of the task as he commented, 'This is not rocket science. It's much harder than that.'

References

Chi, M.T.H., Feltovich, P. & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, **5**, 121–152.

Handy, C. (1994). *The empty raincoat*. London: Hutchinson.

IAEA Conference 2008: papers and presentations. http://www.iaea2008.cambridgeassessment.org.uk/ca/

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, **23**, 2, 13–23.

Salthouse, T. A. (1991). Expertise as the circumvention of human processing limitations. In: K A. Ericsson & J Smith (Eds), *Toward a General Theory of Expertise: Prospects and Limits*. pp.286–300. Cambridge: Cambridge University Press.