BOOK REVIEW

Validity in educational and psychological assessment, by Paul E. Newton & Stuart D. Shaw, London UK, Sage, 2014, 253pp., £25.99 (paperback), ISBN13: 978-1446253236

What makes for a high-quality educational or psychological test? The standard answer is that it must have good reliability and validity (e.g., Clausen-May, 2001). Reliability, the degree to which a test generates consistent results, can be fairly straightforward to document, but validity is much more complicated, as well as being of far greater importance (Black & Wiliam, 2014). Unless a test captures something that we are interested in concerning those who take it, any reliability it may have is of little value. So it is natural to see validity as the supreme criterion for evaluating the quality of an assessment, and hence the importance of the topic addressed by this helpful and timely book.

In this book, Newton and Shaw bring their considerable knowledge and experience to bear on the highly-contested nature of validity in assessments. The relevant literature is huge and can be intimidating for the non-expert, who may find important pieces of writing impenetrable. The authors see the book not as a "how to" guide to validation but as more of a "how to understand" guide to validity and validation. They assume little prior knowledge and take a broadly historical approach, painstakingly charting how conceptions of validity have developed from the mid-1800s to the present day. It is interesting to note that their account features few advances in *techniques* for validation – most of the procedures that we rely on today existed in some form before the 1950s. Instead, the authors' focus is on the changes in thinking concerning what it means to claim validity and how a validity claim may be substantiated.

The book is written very clearly and the reader is led with great care through the details of the story. The authors do not shy away from the complexities of the debates encountered, commenting that "anyone who wants to grapple with validity theory must grapple with some extremely thorny issues in the philosophies of science, measurement and mind" (p. 164). This makes the book at times a demanding read, but very worthwhile and something which everyone serious about assessment should engage with.

The validity of a test is usually defined as the extent to which it measures what it is supposed to measure. To take a simple example, a reading test printed in a very small font size would be likely to be measuring eyesight to some extent rather than solely the literacy that it was constructed to assess. It would be possible for someone to fail the *test situation* rather than the intended *test subject* of reading, and this would therefore constitute an example of weak validity. However, this simple example immediately raises an issue returned to throughout the book – that validity depends on the *purpose* for which a test is constructed or used. Optometrists *do* use reading of text in small font sizes to assess eyesight, so this same test might have good validity if used for a different purpose. For the optometrist, the test would

have poor validity for someone who was not a fluent reader. So it seems to make more sense to see validity as a property of an *interpretation* or a *use* of a test, rather than of the test itself.

Validity of assessments has been a much discussed issue, and over the years there has been a huge proliferation in the different varieties of validity suggested and debated. The authors provide a table (p. 8) containing 151 different kinds of validity that have been proposed since the 1940s, many of which are synonymous, or nearly so, and they comment that various other kinds still continue to be offered today. (Academics can sometimes be good at renaming minor variations of familiar concepts in ways sufficiently obscure to avoid accusations of plagiarism!) It is to the authors' credit that they avoid getting bogged down in the minutiae of this and manage instead to lead the reader through the entire book mainly by reference to just three categories of validity: content validity, criterion validity and construct validity. *Content validity* concerns the inclusion within a test of all relevant aspects of a construct (and may be seen as a more rigorous version of *face validity*). *Criterion validity* concerns how the measure compares with other measures, and is often further divided into *predictive* validity (where the comparison is with something that is assessed at a later time) and *concurrent* validity (where the comparison is with something that is assessed at the same time). Finally, *construct validity* concerns the extent to which a test measures the intended construct. The non-expert might well think that this is just another way of saying that it measures what it is supposed to measure, and this is indeed a position taken by some (e.g., Bramley, 2012). Certainly, of these three kinds of validity, construct validity is the most disputed and hardest to pin down, as it relates to what theories say about what a measure would be expected to look like.

To a large extent, the story of this book is a tale of three validities. The authors recount how the concept of validity was fragmented into content validity, criterion validity and construct validity – the so-called trinitarian conception of validity – and then how one of these parts (construct validity) swallowed up the other two so that everything became construct validity. Because of this convergence into construct validity, there can be confusion now when someone speaks of construct validity, as it may not be clear whether they are referring to it as one part of the trinitarian package or as the generalised version encompassing all of validity.

After charting the early years from the mid-1800s, the authors describe the two basic approaches to establishing validity that were prevalent in the early twentieth century: logical analysis of test content (a forerunner of content validity) and empirical correlation evidence (which would later be classed as criterion validity). Lee Cronbach (famous for being the originator of Cronbach's alpha) described these different forms of validity and drew a clear distinction between measurement and prediction, taking the view that for the purposes of prediction it did not matter what a test measured provided that it predicted something well enough for some practical purpose. He was primarily interested in the process of validation rather than the theory of validity, and it was he, together with Paul Meehl, who most famously elaborated the notion of construct validity. In this way, from the mid-1950s the traditional trinitarian conception of validity as *content validity – criterion validity – construct validity* began to emerge.

The authors then relate how in the mid-1970s, and principally as a result of the work of Samuel Messick, it began to be increasingly acknowledged that content validity and criterion validity were over-simplistic and by themselves insufficient. Verifying that the content of a test adequately sampled the content of its domain said nothing about whether it would elicit the responses that it should from those being tested. Similarly, for criterion validity, a good correlation between a predictor measure and a criterion measure did not mean that we necessarily knew what, if anything, the predictor was measuring. The authors report how Raymond Cattell had been highly dismissive of validating tests by correlation with previous tests, a practice which he had referred to as "holy apostolic succession" (p. 49). Cronbach's early view (c. 1949), as we have seen, was that this did not matter if the predictor predicted well, but this was less true of his later writings. In this respect, Cronbach was a forerunner of Messick, who "dismissed the legitimacy of blindly empiricist approaches to test development" (p. 107). Messick presented his ideas as a *progressive matrix*, consisting of four cells, relating to the test's interpretation and its use, on the one hand, and the evidential and consequential bases on the other. Messick included ethics at the centre of validity theory, with a concern for the human consequences of tests, seeing such *consequential validity* as a central part of the process. But his progressive matrix was difficult to understand and the authors regard this work as confusing and ultimately a mistake. However, over time, under Messick's influence, the trinitarian view of validity gave way to a unitarian perspective, as content validity and criterion validity became subsumed within construct validity as the single overarching kind of validity. In this way, construct validity moved from being the last resort, when neither content validation nor criterion validation was adequate, to becoming the overriding category of validity. By 1999 and the fifth edition of the authoritative *Standards* (AERA et al., 1999), there was a huge consensus in favour of Messick's unified conception of validity.

Bringing the story up to the present, the authors argue that in a sense we are "all construct validity theorists now" (p. 171), although that does not mean that the disagreements have ceased. Since 1999, scholars have increasingly questioned whether the notion of construct validity can bear the weight being placed on it, and various attempts have been made to simplify the approach. In particular, where Cronbach and Messick might be characterised as idealistic, Michael Kane's approach is pragmatic. Kane takes an argument-based approach to validation, seeing it as a rhetorical process which he has likened to building a bridge over a river from the test performance on one bank to the claim that it represents a particular level of some attribute on the other (p. 139).

The authors see "no signs that debate over the scope of validity is likely to be resolved one way or the other in the near future" (p. 178). They conclude the book by focusing on the relationships between the various dimensions of evaluation of testing policy and by providing a new framework for this, which they present as a reinterpretation of Messick's original progressive matrix. Their reworking consists of seven interdependent questions covering both technical evaluation and social evaluation across measurement, decision-making and policy objectives. For example, one of their questions is: "Is it acceptable to implement (or continue

implementing) the testing policy?" (p. 186). They avoid labelling any of their cells as "validity", preferring instead to see the framework holistically. They offer a detailed explanation and defence of the details of their proposal, and it remains to be seen in what ways it will be taken up and adopted and adapted further.

Approaching this fascinating account of the discussions and debates in validity and validation from the current mathematics education context is sobering. It is well-known that poorly-conceived high-stakes tests can have a devastating backwash effect on teaching and learning in classrooms (Berliner, 2011). Teachers are under immense pressure to prioritise classroom activities that match closely the kinds of items that will be in the test, and this raises a serious issue of consequential validity. As the working party of the International Society for Design and Development in Education (Black et al., 2012) put it:

> Driven by pressures for low cost, simplicity of grading and task predictability, current tests have a narrow range of item types that does not reflect the breadth of world-class learning goals as set out, for example, in the Common Core State Standards for Mathematics ... Good tests combine valid and reliable information for accountability purposes with a beneficial influence on teaching and learning in classrooms – i.e. they are tests worth teaching to.

Black et al. suggested that the key to validity is to "assess the things that you are really interested in" – for example, solving substantial mathematical problems – rather than things that might appear easier to assess reliably. Daro and Burkhardt (2012) advocated psychometric test design by taking a structured sample from a population consisting of a large bank of rich mathematical tasks which could be ranked in difficulty by trialling. They suggested publishing the population of tasks along with the balancing algorithm for constructing tests from it.

Currently, with the predictable routine items which tend to dominate most high-stakes assessments, "teaching to the test" may largely undermine claims to validity (Foster, 2013). The US National Research Council (2001) highlighted the compromises that arise when a single assessment is used for multiple purposes and complained that the aspects of learning targeted by large-scale assessments are often out of line with those advocated for use in the classroom. They commented that much more time, money and effort goes into large-scale assessments of individuals and programmes than into classroom assessment designed to assist learning and urged that "investment must be shifted toward the classroom, where teaching and learning occur" (p. 9). Investing in high-quality assessment, which samples from a more authentic problem-solving content domain, could potentially be a lever to elevate the experiences that students are offered in mathematics classrooms. In a context of escalating emphasis on high-stakes assessment, it is crucial that the implications of assessment are carefully weighed up, so that assessment becomes a way to support positive change in students' learning of mathematics rather than something that schools and teachers have to continually battle against.

References

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME)(1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*(3), 287-302.

Black, P. & Wiliam, D. (2014) Assessment and the design of educational materials. *Educational Designer*, *2*(7). Retrieved from: http://www.educationaldesigner.org/ed/volume2/issue7/article23/

Black, P., Burkhardt, H., Daro, P., Jones, I., Lappan, G., Pead, D., & Stephens, M. (2012). High-stakes examinations to support policy: Design, development and implementation. Report of an ISDDE Working Group. *Educational Designer*, *2*(5). Retrieved from: http://www.educationaldesigner.org/ed/volume2/issue5/article16/

Bramley, T. (2012). "Measurement" and "construct" need to be clarified first: Commentary on Newton, P. E. "Clarifying the consensus definition of validity". *Measurement: Interdisciplinary Research & Perspective*, *10*(1-2), 42-45.

Clausen-May, T. (2001). *An Approach to Test Development*. Slough: NFER.

Daro, P. & Burkhardt, H. (2012). A population of assessment tasks. *Journal of Mathematics Education at Teachers College*, *3*(1), 19-25.

Foster, C. (2013). Resisting reductionism in mathematics pedagogy. *Curriculum Journal, 24*(4), 563–585.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. Pelligrino, J., Chudowsky, N., & Glaser, R., (Eds). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Colin Foster

School of Education, University of Nottingham, Nottingham, UK

Email: colin.foster@nottingham.ac.uk