Neurocinema (2011). *Rise of Neurocinema: How Hollywood Studios harness your brainwaves to win Oscars*. Available online at: http://www.fastcompany.com/1731055/rise-neurocinema-how-hollywood-studios-harness-your-brainwaves-win-oscars (Accessed 11 December 2013).

Neurorelay (2012). *Insights from "The Buying Brain: Secrets for Selling to the Subconscious Mind" Book Review*. Available online at: http://neurorelay.com/2012/05/17/insights-from-the-buying-brain-secrets-for-selling-to-the-subconscious-mind-book-review/. (Accessed 13 March 2013).

NIH (2012). *NIH Toolbox Brochure*. Available online at: http://www.nihtoolbox.org/WhatAndWhy/Assessments/NIH%20Toolbox%20Brochure-2012.pdf. (Accessed 16 April 2013).

New York Times (2012). *Academic 'Dream Team' Helped Obama's Effort*. Available online at: http://www.nytimes.com/2012/11/13/health/dream-team-of-behavioral-scientists-advised-obama-campaign.html. (Accessed 25 March 2013).

New York Times (2013). *Obama Seeking to Boost Study of Human Brain*. Available online at: http://www.nytimes.com/2013/02/18/science/project-seeks-to-build-map-of-human-brain.html?pagewanted=all&_r=0. (Accessed 25 March 2013).

OECD (2002). *Understanding the Brain – Towards a New Learning Science*. OECD, Paris.

OECD (2007). *Understanding the Brain: the Birth of a Learning Science*. OECD, Paris.

Prism (2013). *Prism Brain Mapping: Using Neuroscience to improve performance*. Available online at: http://www.prismbrainmapping.com/ (Accessed 30 January 2013).

Räsänen, P., Salminen, J., Wilson, A.J., Aunio, P. and Dehaene, S. (2009) Computer-assisted intervention for children with low numeracy skills. *Cognitive Development*, **24**, 4, 450–472. Available online at: http://www.aboutdyscalculia.org/author.html (Accessed 9 July 2013).

Rust, J. & Golombok, S. (1999) 2nd edition *Modern Psychometrics: The Science of Psychological Assessment*, Routledge, London & New York.

Schreiber, D. (2007). Political Cognition as Social Cognition: Are we all political sophisticates? In: Neuman, W.R., Marcus, G. E., Crigler A.N., *et al*. (Eds) *The Affect Effect: Dynamics of Emotion in Political Thinking and Behavior*. Chicago, IL: University of Chicago Press, pp.48–70.

Stevenson Jr., A.E. (1952) *Speech at the University of Wisconsin*. Madison, 8 October 1952.

Szücs, D. & Goswami, U. (2007). Educational neuroscience: Defining a New Discipline for the Study of Mental Representations. *Mind, Brain and Education*, **1**(3), 114–127.

Szücs, D., Devine, A., Soltesz, F., Nobes, A., Gabriel, F. (2013). Developmental dyscalculia is related to visuo-spatial memory and inhibition impairment. *Cortex*. Available online at: http://www.sciencedirect.com/science/article/pii/S0010945213001688#. (Accessed 26 September 2013).

TES (2013). *Get inside their heads*. TES, 1 March 2013, pp.28–32.

Todd, R.L., Berninger, V. W., Stock, P., Altemeier, L., Trivedi, P. & Maravilla, K. R. (2011). Differences between good and poor child writers on fMRI contrasts for writing newly taught and highly practiced letter forms. *Reading and Writing*, **24**, 5, 493-516. Available online at: http://link.springer.com/article/10.1007%2Fs11145-009-9217-3 (Accessed 12 March 2013).

Westen, D. (2007). *The political brain: The role of emotion in deciding the fate of the nation*. New York. Public Affairs Books.

Wilson, A.J., Revkin, SK., Cohen, D., Cohen, L. & Dehaene, S. (2006). An open trial assessment of The Number Race, an adaptive computer game for remediation of dyscalculia. *Behavioural and Brain Functions*, **30**, 2, 20.

Book announcement: Validity in Educational and Psychological Assessment

Paul Newton Institute of Education, University of London (formerly Cambridge Assessment) and **Stuart Shaw** Cambridge International Examinations

Introduction

For almost one hundred years, divergent views on the concept of validity have proliferated. Even today, the meaning of validity is heavily contested. Despite a century of accumulated scholarship, new definitions of validity continue to be proposed, and new 'types' of validity continue to be invented (see Newton and Shaw, 2013). Yet, against the backdrop of an evolving measurement and testing landscape and the increased use of assessments across scientific, social, psychological and educational settings, validity has remained "the paramount concept in the field of testing." (Fast and Hebbler, 2004, p.i).

Validity is universally regarded as the hallmark of quality for educational and psychological measurement. But what does quality mean in this context? And to what exactly does the concept of validity actually apply? What does it mean to claim validity? And how can a claim to validity be substantiated? In a book entitled *Validity in Educational and Psychological Assessment*, due to be published in the UK by SAGE in March 2014, we explore answers to these fundamental questions.

Validity in Educational and Psychological Assessment adopts an historical perspective, providing a narrative through which to understand the evolution of validity theory from the nineteenth century to the present day. We describe the history of validity in five broad phases, mapped to the periods between:

1. the mid-1800s and 1920: gestation
2. 1921 and 1951: crystallisation
3. 1952 and 1974: fragmentation
4. 1975 and 1999: (re)unification
5. 2000 and 2012: deconstruction.

We explain how each of these phases can be characterised by different answers to the question at the heart of any validation exercise: how much and what kind of evidence and analysis is required to substantiate a claim of validity?

The book comprises six chapters. In Chapter 1 we set the scene for the historical account which follows. Chapters 2 to 5 offer readers a chronological account that delineates the phases of development of validity theory and validation practice. In Chapter 6 we propose a framework for the evaluation of testing policy, which we based on the original progressive matrix from Messick (1980).

Chapter 1: Validity and Validation

In Chapter 1 we begin by exploring a range of everyday and technical meanings of validity in order to set the scene for the historical account which follows. This is an account of validity as a technical term of educational and psychological measurement, which is important to bear in mind because the term 'validity' has very many different meanings, some of which are entirely independent of measurement. The main chapters of the book attempt to demonstrate how, even within this relatively narrow conceptualisation, its meaning is still nevertheless contested and resistant to precise definition. Yet it needs to be appreciated, from the outset, that it does mean something quite distinctive in this particular context, even if that 'something' might be difficult to articulate.

Following a discussion of the conventions used in the textbook we present an outline of the history of validity. The historical account is our attempt to describe and to explain how conceptions of validity and validation have evolved within the field of educational and psychological measurement.

Our historical account tends to focus more on concepts of validity theory than on the practice of validation. Good validation practice is the application of good validity theory. In the absence of validity theory there is nothing to guide or to defend validation practice. It is theory that constitutes the rational basis for validation practice. As we discuss each new contribution to the theory of validity, their implications in terms of a positive, operational impact upon validation practice become increasingly apparent.

Chapter 2: The Genesis of Validity (mid-1800s to 1951)

Chapter 2 covers the first two phases outlined above: a gestational period, from the mid-1800s to 1920; and a period of crystallisation, from 1921 to 1951. The chapter is heavily skewed towards the latter, as the period during which the concept of validity developed an explicit identity or, perhaps more correctly, a range of different identities.

In this chapter, we explore early conceptions of validity and validation, focusing particularly upon achievement tests, general intelligence tests, and special aptitude tests. We argue that the emergence of validity as a formal concept of educational and psychological measurement can only be understood in the context of major developments in testing for educational, clinical, occupational and experimental purposes which occurred during the second half of the nineteenth century and the early decades of the twentieth century, most notably in England, Germany, France and the USA. Upon this foundation was proposed the 'classic' definition of validity: the degree to which a test measures what it is supposed to measure.

Although there are numerous accounts of the history of validity theory and validation practice during the early years (e.g. Anastasi, 1950; Geisinger, 1992; Shepard, 1993; Kane, 2001) the impression given is often of a period almost exclusively dominated by prediction, the empirical approach to validation, and the validity coefficient. Reflecting on this period, Cronbach (1971) observed that the theory of prediction was very nearly the whole of validity theory until about 1950; a characterisation later endorsed by Brennan (2006). Kane (2001) characterised the early years as the 'criterion' phase, where the criterion was typically understood as the thing that was to be predicted.

The impression given by a number of notable chroniclers (e.g. Moss, Girard and Haniford, 2006) is that the key developments in validity theory can be traced either to successive editions of *Educational Measurement*, beginning with Lindquist (1951) or to successive editions of professional standards documents, beginning with American Psychological Association/American Educational Research Association/National Council on Measurements Used in Education (APA, AERA, NCMUE, 1954). We argue that there is a far more interesting story to be told about the early years. We contend that many of the developments in validity theory and validation practice, from the middle of the twentieth century onwards, are simply elaborations of earlier insights. The earliest definition of validity was far more sophisticated than the idea of a validity coefficient might suggest, and the earliest approaches to validation were far more complex and involved. Education took a lead in formally defining the concept, and achievement testers, aptitude testers, intelligence testers and personality testers played their role in refining it and developing new techniques for investigating it.

The more interesting story of validity during the early years is one of sophistication and diversity; at least in terms of ideas, if not always in terms of practice. Because of its diversity, though, it is hard to characterise the period succinctly.

Chapter 3: The Fragmentation of Validity: 1952 to 1974

The diversity of ideas on validity and validation during the early years presented a challenge to test developers and publishers. Given a variety of approaches to validation to choose from, and with even the experts valuing those approaches quite differently, how were professionals in the field to decide what information on test quality they needed to make available to consumers? And, in the absence of agreement upon principles of best practice and specific guidelines about criteria for the evaluation of tests and testing practices, how were test developers and publishers to be held to account?

The first edition of what was to become known as the *Standards* (APA, AERA, NCMUE, 1954) was written to make sense of the landscape of the early years. As a consensus statement of the professions, the *Standards* included both implicit standards for thinking about validity and explicit standards for conducting and reporting validation research.

The *Standards* emphasised 'types' of validity, specialised to the contexts of test use: content validity, predictive validity, concurrent validity, and construct validity. If, for example, you needed to validate an interpretation drawn in terms of achievement, then you needed to adopt a particular approach to validation, content validation, which meant establishing a particular kind of validity, content validity. Although these were explicitly described as "Four types of validity" (APA, AERA, NCMUE, 1954, p.13) the *Standards* was a little confused over the matter, also describing them as 'aspects' of a broader conception.

Between 1954 and 1974, the *Standards* was revised twice, in order to respond to constructive criticism, to take account of progress in the science and practice of educational and psychological measurement, and to respond to societal change. Yet, mixed messages continued to be promulgated over the nature of validity. For many who were influenced by the *Standards* during this time, they came to embody and to cement a fragmented view of validity and validation, whereby different uses to which test scores were to be put implied different approaches to validation and even different kinds of validity.

Chapter 4: The (Re)Unification of Validity: 1975 to 1999

Samuel Messick's account of validity and validation became the zeitgeist of late twentieth century thinking on validity during the 1980s and 1990s. Developing ideas from Harold Gulliksen and Jane Loevinger, and with the support of allies including Robert Guion, he brought the majority of measurement professionals of his generation around to the viewpoint that all validity ought to be understood as construct validity. His thesis was that measurement ought to be understood (once more) as the foundation for all validity; and therefore that construct validation – scientific inquiry into score meaning – ought to be understood as the foundation for all validation.

Through an extended discussion of Messick's contribution to validity theory, we describe this period in terms of his *triumph* and his *tribulation*. Messick was enormously successful in promoting validity as a unitary concept, in contrast to earlier fragmented accounts. His triumph, therefore, concerned the *science* of validity: he convinced the educational and psychological measurement communities that measurement-based decision-making procedures (i.e. tests) needed to be evaluated holistically, on the basis of a scientific evaluation into score meaning. Enormously problematic, though, was his attempt to integrate values and consequences within validity theory through his famous (if not infamous) progressive matrix. Unfortunately, not only was his account confusing, it also seemed a little confused. His tribulation, it seems fair to conclude, concerned the *ethics* of validity. Messick failed to provide a convincing account of how ethical and scientific evaluation could straightforwardly be integrated.

In retrospect, it seems hard to disagree with the conclusion, drawn by Shepard (1997), that Messick's progressive matrix was a mistake. Having said that, we believe that its underlying intention was an excellent one. It was an attempt to emphasise that the following two questions were both crucial to any thorough evaluation and were inherently interrelated:

1. Is the test any good as a measure of the characteristic it purports to assess?
2. Should the test be used for its present purpose?

Messick's progressive matrix was supposed to explain the relationship between these two questions, and their relation to the concept of validity, but it was muddled. As Messick helped readers to find their way through the ambiguity of the matrix, his presentation became clearer, but also narrower, as scientific questions of test score meaning began to gain prominence while ethical questions of test score use were nudged into the wings.

Unfortunately, Messick's tribulation led to one of the most notorious debates of all time concerning the scope of validity theory. The field is now genuinely split as to whether, and if so how, evidence from consequences ought to be considered part of validity theory - an issue we tackle in Chapter 5.

Chapter 5: The Deconstruction of Validity (2000 to 2012)

During the 1990s, work on validity and validation was heavily influenced by Messick. The fifth edition of the *Standards* (American Educational Research Association/ American Psychological Association/National Council on Measurement in Education, 1999) was essentially a consensus interpretation of his position, that is, a unified conception of validity. The *Standards* reflected the prevailing view of the time - a construct-centred approach to validity. Yet, with the turn of the millennium, cracks began to emerge. On one hand, it was unclear how to translate construct validity theory into validation practice. On the other hand, it was unclear whether construct validity was actually the best way to unify validity theory. It seemed that an element of deconstruction might be in order, reflecting the desire to simplify validation practice as well as the desire to simplify validity theory.

In terms of validation practice, this period was characterised by growing consensus over the value of a new methodology for guiding, and simplifying, validation practice. Argumentation, it now seemed, held the key. Michael Kane had developed a methodology to support validation practice, grounded in argumentation (e.g. Kane, 1992). This provided a framework, or scaffold, for constructing and defending validity claims. Thus, while Messick defined the claim to validity in terms of an overall evaluative judgement, Kane explained exactly how that claim to validity could be constructed and defended. The argument-based approach took a long time to take root, though, and only began to have a significant impact well into the new millennium. In fact, even having begun to take root, it still proved surprisingly challenging to implement. Goldstein and Behuniak (2011) noted that very few examples are available to the research community of validity arguments for large-scale educational assessments.

In terms of validity theory, this period was characterised by growing controversy, embodied in two major debates. The first concerned the nature and significance of construct validity: a debate over the relatively narrow, scientific issue of score meaning. A critical question was whether construct validity ought to be considered the foundation of all validity, as Messick had argued. Related questions concerned whether all validation needed to be understood in terms of constructs; whether the nomological networks of Cronbach and Meehl (1955) were useful or even relevant to validation; whether validity was a concept more like truth or more like justified belief; whether validity ought to be theorised in terms of measurement; and whether the concept of validity could be applied in the absence of standardised procedures.

The second concerned the scope of validity: a debate over whether the concept ought to be expanded beyond the relatively narrow, scientific issue of score meaning, to embrace broader ethical issues concerning the consequences of testing. Various 'camps' developed: from liberals, who extended the use of 'validity' to embrace social considerations of test score use; to conservatives, who restricted the use of 'validity' to technical considerations of test score meaning.

Chapter 6: 21st Century Evaluation

The concept of validity has assumed a pivotal role across decades of debate on the characteristics of quality in educational and psychological measurement. Despite this, it has proved extremely resistant to definition. In Chapter 6, we respond to the concerns of the more conservatively minded, who object that the concept of validity is becoming so large as to present an obstacle to validation practice. We do so by proposing a new framework for the evaluation of testing policy. In fact, we see this as a revision of the original progressive matrix from Messick (1980), which we have redesigned to dispel some of the confusion engendered by its original presentation. After first defending the new framework we then provide a more detailed analysis of technical and social evaluation, before considering evaluation within each of the cells respectively.

Validity in Educational and Psychological Assessment will be available from March 2014. The authors believe that this book will be of interest to anyone with a professional or academic interest in evaluating the quality of educational or psychological assessments, measurements and diagnoses.

ISBNs: Paperback: 9781446253236 Hardback: 9781446253229

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, **51**, 2, Supplement.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational & Psychological Measurement*, **10**, 1, 67–78.

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Westport, CT: Praeger Publishers.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.). *Educational Measurement* (2nd Edition) (pp.443–507). American Council on Education. Washington: DC.

Cronbach, L.J. & Meehl, P.E (1955) Construct validity in psychological tests. *Psychological Bulletin*, **52**, 4, 281–302.

Fast, E.F. & Hebbler, S. with ASR-CAS Joint Study Group on Validity in Accountability Systems. (2004). *A Framework for Examining Validity in State Accountability Systems*. Washington, DC: Council of Chief State School Officers.

Geisinger, K.F. (1992). The metamorphosis of test validation. *Educational Psychologist*, **27**, 2, 197–222.

Goldstein, J. & Behuniak, P. (2011). Assumptions in alternate assessment: An argument-based approach to validation. *Assessment for Effective Intervention*, **36**, 3, 179–191.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, **112**, 3, 527–535.

Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, **38**, 4, 319–342.

Lindquist, E.F. (Ed.) (1951). *Educational Measurement*. Washington, DC: American Council on Education.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, **35**, 11, 1012–1027.

Moss, P. A., Girard, B. J & Haniford, L. C. (2006). Validity in Educational Assessment. *Review of Research in Education*, **30**, 1, 109–162.

Newton, P.E. & Shaw, S.D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, **18**, 3, 301–319.

Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education*, **19**, 405–450.

Shepard, L.A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, **16**, 2, 5–8.

Research News

Karen Barden Research Division

Conferences and seminars

The Future of Education International Conference

In June, Sanjana Mehta attended The Future of Education Conference in Florence, Italy. The conference aims to promote transnational cooperation and share good practice in the field of innovation for education. Sanjana presented a paper on *Thrown in at the deep end? Exploring students', lecturers' and teachers' views on additional support lessons at university.*

The Assessment in Higher Education Conference

Held in Birmingham in June, this fourth biennial conference provided an opportunity to debate key issues and developments in current assessment, policy and practice. Simon Child presented a paper entitled *"I've never done one of these before". A comparison of the assessment 'diet' at A level and the first year of university.*

British Education Studies Association (BESA)

The ninth BESA Annual Conference took place at Swansea Metropolitan University in June. The key theme of the conference was Education: Past, Present and Future. Jackie Greatorex presented on *Using scales of cognitive demand in a validation study of Cambridge International A and AS level Economics*.

Journal of Vocational Education and Training (JVET)

The JVET tenth international conference was held in July at Worcester College, Oxford. Colleagues from the Research Division presented the following papers:

Jackie Greatorex: *How can major research findings about returns to qualifications illuminate the comparability of qualifications?*

Martin Johnson: *Insights into contextualised learning: how does feedback on performance contribute to professional examiners' shared understanding?*