CAIE 1073479
25 July 2015
Initial
CE: SV QA: MA
Coll: QC:

Assessment in Education: Principles, Policy & Practice, 2015

Routledge
Taylor & Francis Group

REVIEW ESSAY

Zen and the art of validity theory

Validity in educational and psychological assessment, by P.E. Newton & S. Shaw, London, Sage, 2014, 280 pp., £25.99 (paperback), ISBN: 978-1-4462-5323-6

In *Validity in Educational and Psychological Assessment*, Paul Newton and Stuart Shaw provide a unique historical and conceptual overview of the psychometric concept of validity, as it arises in psychological and educational testing. In addition, they propose a novel integrated account of questions as they relate to validity in testing situations. Thus, the book offers invaluable historical material, and a conceptual framework that will help readers tease apart the many issues relevant to validity theory. In my view, that's a bargain, so if you're in a hurry and not inclined to read the rest of this review, let me wrap it up by giving you a simple piece of advice: if you are even remotely interested in validity theory, buy this book.

Content overview

To the best of my knowledge, this is the first time scholars have attempted to provide a more or less complete study into the historical development of the validity concept. The book is structured in terms of four historical periods that Newton and Shaw identify in terms of conceptual movements: the genesis of validity theory (mid 1800s–1951), the fragmentation of validity (1952–1974), the (re)unification of validity (1975–1999) and the deconstruction of validity (2000–2012). Each of these periods has its own chapter, which largely consists of a factual overview of the work appearing in the relevant period together with a substantive interpretation of that work in terms of the prevailing scientific, philosophical and sociocultural winds of the time.

In *the genesis of validity theory*, Newton and Shaw have dug up many treasures. They have, among other things, located the earliest references to validity that I have come across. In addition, the early days of validity theory, as Newton and Shaw sketch them in the opening chapter, prove to be surprisingly rich in ideas and concepts, and feature a level of scholarly creativity and originality that makes some of the later periods pale in comparison. The early scholars of psychometrics keenly and quickly identified the important matters in psychological testing: validity, which was taken to be the question of whether the test measures what it purports to measure, and reliability, which referred to the stability of measurement outcomes in cases where the measured attribute was constant. On the topic of validity, the late nineteenth century and early twentieth century saw a great pluralism of opinions that, despite their old age, strangely struck me as fresh. Newton and Shaw give a fascinating overview of the depth and variety of validity theory in these days, and for me that was one of the highlights of the book.

In *The fragmentation of validity*, Newton and Shaw argue that the philosophical movements of the time (operationalism, positivism and behaviourism) led validity theory to identify the concept of validity with its operationalization. This came down to a reduction of validity to either criterion validity (how well does the test score predict behaviour outside of the testing situation?) or content validity (how well does the test's content match, e.g. an educational curriculum?). Neither of these concepts went quite to the core of validity because, as Newton and Shaw suggest, the distinction between these types of validity is a distinction between kinds of evidence for the hypothesis of validity rather than a distinction between genuinely separate validities. In 1955, this issue came to the fore with the publication of Cronbach and Meehl's seminal paper introducing the notion of construct validity. Somewhat half-heartedly (I personally think this was a strategic move rather than a sincere opinion of the authors), their paper introduced construct validity as if it only applied to situations in which criterion or content validity were not applicable (e.g. because there was neither a clear criterion to be defined, nor a demarcated description of relevant item content, as in personality testing). As a result, construct validity was originally *juxtaposed* with criterion and content validity, which led to the so-called trinitarian view of validity.

But of course that could not last. Construct validity was a cuckoo's egg among the other types of validity, especially because, in the second half of the twentieth century, everyone had somehow become convinced that *all* test scores were supposed to measure theoretical constructs (I personally think this is one of the capital mistakes in twentieth-century test theory). As a result, the trinitarian view became conceptually unstable, and the notion of construct validity took its place as the new centrepiece of validity theory.

This led to an era that Newton and Shaw label *The (re)unification of validity*: a period that is characterised by attempts to bring the different kinds of validity under one header. In this period, construct validity became the one kind of validity, which was to be considered in every situation where test score interpretation and use was in need of justification. It was also during this period that construct validity theory took its characteristic form or, on a less positive reading, acquired some of its current dogmatic shape. For instance, the idea took hold that validity applies to interpretations of test scores rather than to tests, the methodological community became wedded to the idea that validity research is never-ending or at least a superhuman feat, and theorists started producing Zen-like phrases such as 'all construct validity is one' (Cronbach, 1980) that became quite prevalent in the validity literature. In addition, during this time, the consequences of testing started to play a role in the evaluation of validity of testing programmes. Cronbach initiated this movement in his 1971 chapter in *Educational Measurement*, and Samuel Messick completed it in his 1989 chapter in the same book which, without doubt, is the landmark event of the (re)unification period.

Although, as Newton and Shaw convincingly argue, Messick's chapter bordered on inconsistency, and has proven very hard to understand for most people, it nevertheless became the blueprint for the 1999 *Standards*. The great synthesis of validity that Messick (1989) sought to produce, and its 1999 APA-certified copy, certainly carried an enormous conceptual weight. The theory required a consideration of just about every fact about a test or testing procedure that one could imagine was remotely relevant to validity, and an integration of these pieces of information into what Messick called an 'overall judgement'. As there was no guide on how to do the

CAIE 1073479
25 July 2015
Initial
CE: SV QA: MA
Coll: QC:

required integration of information, this advice left many researchers behind in a state of bewilderment.

Messick in addition defined the judgment itself (not the property being judged) as validity, something that many have felt to be too large a departure from the core meaning of validity. Finally, Messick's validity theory continuously, and without exception, stressed the importance of theoretical constructs – even in very low-level cases of testing where talk of constructs appears to be superfluous. Newton and Shaw offer an extended criticism of Messick's theory, although in their final chapter they build on aspects of his view (most notably his famous progressive matrix) to construct their own conception of validity.

As a countermovement, over the past decades, many have left Messick's synthesis behind, and have sought to create more lightweight validity concepts. My own work is a case in point (Borsboom, Mellenbergh, & van Heerden, 2004), but so is Michael Kane's argument-based approach (Kane, 2006), which, in many cases, shuns talk of theoretical constructs and nomological networks. Instead, Kane stresses the need to produce a strong argument for *whatever* test score interpretation and use one intends to defend, and presents this as a technique which can be used in tandem with any methodological, conceptual or philosophical inclination.

Newton and Shaw discuss this period in *The deconstruction of validity*, and do a good job of representing the lively debate on foundational issues in validity theory that has ensued in the past years, covering many of the arguments that have been brought forward, even though some of these, in my view, are given too scant attention (e.g. the relevance of the distinction between formative and reflective measurement models in evaluating validity, and the general issue of whether we should require a causal connection to exist between the measured attribute and the test score; Bollen & Lennox, 1991; Markus and Borsboom, 2013).

The book rounds off with a concluding chapter that both attempts to produce a synthesis of all that has come before, and to sketch a road forward. Newton and Shaw's sketch is basically an extension of Messick's progressive matrix, and, like Messick's validity concept, deals with many different questions that one might ask about a testing procedure at the same time. The proposed framework resembles a set of guidelines for programme evaluation rather than a psychometric theory of validity. As such, I think it will likely prove more attractive to those involved in large-scale testing programmes than to those involved in scientific research or theoretical psychometrics. This is because Newton and Shaw's theory is neither designed nor able to deal with the crucially important scientific and psychometric question of whether a test measures what it should measure; for it does not contain a theory on what measurement is, on when it is achieved or on what it means to measure the attribute designated by a particular theoretical term. Thus, like its predecessors, and like currently popular argument-based accounts, the theory leaves a hole in the heart of psychometrics which is not easily filled.

It is unclear to me whether Newton and Shaw appreciate the significance of this problem, but in my view both scientific methodology and psychometric theory require a validity theory that can lay out the truth conditions for sentences like 'this IQ-test measures general intelligence'. One example of such a theory is my own causal account, which states that test X is valid for attribute Y if and only if variation in attribute Y causes variation in the scores on test X (Borsboom et al., 2004), but of course other accounts are possible (Markus and Borsboom, 2013). However, Newton and Shaw's theory, like almost all accounts of the so-called

consensus variety, is unable to perform this task; thus, ironically, although today's validity literature is designed to answer almost any question that might possibly arise in the use and interpretation of test scores, it cannot answer the very question that got validity theory started: does our test measure what we want it to measure?

With that caveat, however, I do think that Newton and Shaw have succeeded in articulating the three important questions that invariably arise in testing: (1) what, if anything, are we picking up with our test scores, (2) how good is the evidential argument to back up the intended interpretation of test scores and (3) given a set of goals, political convictions and ethical standards, should we implement the testing procedure or not? Interestingly, in this respect Newton and Shaw's subdivision of the issues involved in testing align virtually perfectly with the central principles Keith Markus and I defined in our recent book[1] (Markus and Borsboom, 2013, Ch 12).

Criticisms and omissions

Naturally, there are several points at which I disagree with Newton and Shaw, and despite the wide coverage of the book there are still some important omissions. Perhaps, the most conspicuously missing topic in this book is the test itself. Newton and Shaw give virtually no consideration to the nature and content of the psychological tests that the validity debates are about. Thus, the book has very few discussions of real-world examples. As a result, some discussions become rather ethereal. For example, Newton and Shaw devote attention to the question of whether psychological tests can be concluded to literally *measure* attributes, in view of Michell's (1997) argument to the contrary. But because they do not explain how the measurement process pans out in quantitative measurement in the natural sciences, and neither says what it is that psychological tests miss in this respect, it remains rather unclear to the reader why the question matters at all, or even what it is about. This lack of concrete examples and substantive discussion thereof has made the book much more abstract than it needs to be.

AQ4

Second, Newton and Shaw often analyse the past in terms of the present. That is, they have a habit of interpreting events in, say, the 1950s in terms of theories that only evolved in the 1980s. Thus, there is a lot of 'foreshadowing' in their storyline, and the history that Newton and Shaw wrote is the type of history that 'goes somewhere', i.e. that appears to move toward the current situation in a goal-directed manner. This makes some turns in validity theory, which in my view are best thought of as historical accidents, look like necessary developments and occasionally leads Newton and Shaw to accept explanations of historical events that, in my view, are too shallow.

An example is the emergence of 'construct validity'. Newton and Shaw discuss the struggle of validity theorists with different validity concepts in the fragmentation period and write that 'construct validity still did not present itself as the logical conclusion' (p. 95). This is characteristic of their investigative style: they see construct validity as a concept that would have necessarily presented itself sooner or later, because, well, it is the 'logical conclusion'. Thus, they analyse events as if they 'lead up' to the emergence of construct validity. Because of Newton and Shaw's tendency to analyse history in such an almost teleological way, they do not consider alternative explanations for the historical developments they perceive as necessary. In the case of the dominance of construct validity, for example, one such

explanation could be that a small community of influential people pushed the concept because they believed in it, that some academics gained influence by joining in, and that the concept fitted the interests of certain groups of stakeholders in educational and psychological testing better than precursors to construct validity. Such considerations are unfortunately absent in Newton and Shaw's analysis, which occasionally comes across as somewhat sterile for this reason.

This brings me to a third omission in the book, which is Newton and Shaw's lack of attendance to the question of what kind of field 'validity theory' really is: how large is the field, who dominates it, who influences it? Especially with respect to outside influences, I think this is an opportunity missed. For instance, Newton and Shaw discuss the emergence of the American Psychological Association (APA) as the institutionally dominant force in validity theory through its publication of the *Standards for Psychological and Educational Testing*. But they do not analyse or evaluate that development, and neither do they consider the question of what validity theory might have looked like if the APA had not seized it and brought it under the rule of institutional regulation.

For example, it appears to me that the emergence of a consensus on validity theory towards the 1990s is at least partly a result of the fact that the successive editions of the *Standards* forced it. I find it hard believe that a consensus on validity would have appeared naturally: there are just too many thorny issues involved, and, in comparison to the wider philosophical literature, the consensus that did arise rests on such an idiosyncratic philosophical basis that, at least to me, it has always appeared distinctly unnatural, in the sense of being contrived. Thus, it appears to me that the societal pressure channelled through organisations like APA must have played a major role in the production of the validity consensus. However, whether this analysis is correct or not is not really the point; I merely give it to sketch the kind of reasoning that I missed in the book. In this example, as in several others, Newton and Shaw take an insufficiently wide perspective; that is, they analyse the progression of positions in validity theory as a purely logical or conceptual progression of ideas, and pay little attention to the way in which the content of validity theory was shaped by external forces that have little to do with validity per se.

A final missing topic in the history that Newton and Shaw wrote lies in the personalities of the people who made validity theory into what it is today. It appears to me that validity theory is, and has always been, a field with a very small basis. There are not many people who professionally theorise about validity, which means that the few who do can gain disproportionate influence. This is notably different from, say, psychometric research on reliability and measurement precision, which are squarely embedded in the fields of psychometrics and statistics and to which very many people contribute. In contrast, validity theory was shaped by only a few people, and much of it was actually handcrafted by a single man: Lee Cronbach, whose enormous influence on validity theory is described by Newton and Shaw, but not evaluated. Although I would be the last to deny that Cronbach's work had many virtues, he towered so highly over American methodology for so many years, that no validity theorist could escape his shadow. In such a case, it is historically important to consider the man as well as his ideas. Samuel Messick, whose 1989 chapter on validity receives a much extended discussion in Netwon and Shaw's book, offers a particular example of how far Cronbach's influence went. The story goes that when a senior researcher at ETS remarked to Messick that he found it hard to give feedback on Messick's draft of his 1989 chapter because it was just so

difficult to understand, Messick responded, 'Well, I didn't write it for you. I wrote it for Cronbach!'. I do not have the means to verify this story, but, true or not, it points to a potentially important aspect of the development of validity; for it appears that what wasn't done *by* Cronbach was done *for* Cronbach.

In such a case, one cannot really avoid getting into at least some of the personal and interpersonal details of the development of validity theory. However, such matters are almost entirely neglected by Newton and Shaw. As a result, their history is a history of ideas rather than a history of people. At many points, I do not think this is a problem, because the personalities of people who came up with the ideas discussed don't really matter. But in some cases, like the above, I think Newton and Shaw could have given more attention to the fact that ideas do not spontaneously arise in a vacuum, but are produced by people of flesh and blood; and, where relevant, they might have actually discussed what kind of people they were, who they influenced, and what their function in the academic system was.

Conclusion

Newton and Shaw have written a comprehensive and unique historical overview, which will prove highly useful to current and future scholars in the field. Their account is insightful, clear and honest: the authors do not attempt to hide their sense of wonder, or, in some cases, bewilderment, in discussing the problems that arise in validity theory and the answers that have been given in response to these problems. In addition, the book does not suffer from the immediate need to answer, mitigate or dismiss problems in validity theory: some problems are simply left unanswered and for one working in validity theory, that is a breath of fresh air. Thus, I enjoyed reading this book greatly. Newton and Shaw tell one of the great stories in the history of psychology, and they tell it well. I highly recommend this book.

AQ1

Disclosure statement

AQ6

No potential conflict of interest was reported by the author.

Note

1. The resemblance is so obvious that I feel compelled to state that we worked entirely independently of each other; as Newton and Shaw note in their introduction, they only learned about our book after it was published, and the same holds for us.

References

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305–314.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

AQ7

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

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *New directions for testing and measurement: Measuring achievement over a decade* (pp. 99–108). San Francisco, CA: Jossey-Bass.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355–383.

Denny Borsboom
Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands
d.borsboom@uva.nl

http://dx.doi.org/10.1080/0969594X.2015.1073479