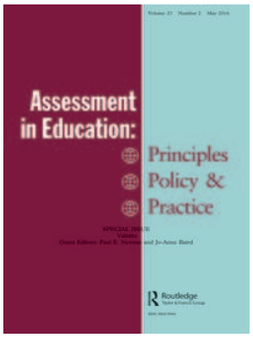

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: http://www.tandfonline.com/loi/caie20

The great validity debate

Paul E. Newton & Jo-Anne Baird

To cite this article: Paul E. Newton & Jo-Anne Baird (2016) The great validity debate, Assessment in Education: Principles, Policy & Practice, 23:2, 173-177, DOI: 10.1080/0969594X.2016.1172871

To link to this article: http://dx.doi.org/10.1080/0969594X.2016.1172871

Assessment in Education: Principles, Policy & Practice, 2016
Vol. 23, No. 2, 173–177, http://dx.doi.org/10.1080/0969594X.2016.1172871

EDITORIAL

The great validity debate

Validity is the most important term in the educational and psychological measurement lexicon. Measurement professionals are generally happy to agree about that. What they are less happy to agree about is what the term ought to mean. North American measurement professionals have negotiated a kind of consensus on this thorny issue, through the definition and description of validity in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, and NCME], 2014). Yet, the status of this consensus is unclear, given continuing debate amongst scholars, and given the fact that all sorts of different definitions and descriptions can be found on the websites of measurement organisations within the USA and elsewhere, and within the pages of prominent textbooks. In short, there is no widespread professional consensus concerning the best way to use the term.

In 1997, Linda Crocker penned an editorial for the North American National Council on Measurement in Education (NCME) publication, *Educational Measurement: Issues and Practice*, entitled: *The Great Validity Debate* (Crocker, 1997, p. 4). Her editorial introduced a special issue of the journal devoted to a controversy which she described as having been 'brewing in psychometric circles' since the late 1980s. It concerned the significance of consequences for the concept of validity and pivoted, for many, around the issue of whether validation should be 'regarded as a scientific, empirical enterprise or a sociopolitical process as well.' She suggested that: 'the prevailing argument in this debate will shape the nature of measurement practice and professional preparation for years to come.'

Well over a decade later, Newton and Shaw undertook an extensive review of the literature on validity, to provide a foundation for an introductory overview of the concept of validity (Newton & Shaw, 2014). Their research led them to conclude that no position in this debate had yet prevailed. Not only was the controversy over consequences still raging (e.g. Cizek, 2012), new controversies had arisen, including debate over the relationship between validity and truth (e.g. Borsboom & Markus, 2013; Borsboom, Mellenbergh, & van Heerden, 2004; Kane, 2013a, 2013b). In an attempt to explore potential for resolving these debates, Newton and Shaw organised a coordinated session at the 2014 NCME Annual Meeting, in Philadelphia, entitled: *What is the Best Way to Use the Term 'Validity'*? The six focal papers at the heart of this new special issue began life in that session. Lorrie Shepard, a contributor to the original special issue (edited by Linda Crocker in 1997) contributed a 'reflective overview' to the session and agreed to provide a similar contribution in this new special issue.

In the spirit of facilitating debate, we decided to introduce an element of peer commentary to the following pages. The six focal papers were prepared simultaneously and then circulated to a group of leading measurement professionals for

comment – on whatever they were inclined to comment on. Once the commentaries had been prepared, they were circulated to the focal paper authors who were invited to respond – again, on whatever they chose to respond to, including both commentaries and focal papers. We hope you will agree that this opportunity for further reflection has paid dividends, in enabling arguments to be honed and positions to be clarified. It has also allowed a larger number of voices to be heard – some whose perspectives have previously been published and others whose have not.

The six focal papers provide an excellent resource through which to understand the various validity debates. It is clear that the controversy over consequences still looms largest. The first and last of these papers (Newton and Shaw and Markus, respectively) reflected on the debate itself – its nature and potential for resolution – whilst the four in between each advocated a different position (Kane, Cizek, Sireci and Moss). Each of these four papers presented views which their authors have expressed before alongside ideas which will be new to readers; thereby sharpening the debate and enabling us to deepen our understanding of the issues at stake. The excellent commentaries and responses speak for themselves.

With such a fundamental concept for assessment under discussion, the debate is philosophical, epistemological, theoretical and methodological. The Editorial Board considered from the outset the danger that the special issue could produce articles in which academics speak only to themselves, with no wider implications being considered; a charge that the Great Validity Debate oftentimes has levelled against it. However, the authors of the focal papers considered the implications of their positions for practice and the range of commentators invited to contribute also intentionally reflected a range of perspectives.

The focal papers, commentaries and responses have shed new light on the many different ways in which measurement professionals agree and disagree over validity and the evaluation of educational and psychological testing. For instance, not everyone agreed with Newton and Shaw that a widespread professional consensus over how best to use the word ‘validity’ would be a good thing. Twing emphatically denied this. Slaney saw the benefit in a reasonable degree of intersubjective agreement, along the lines of a family resemblance definition, but noted that a single consensus definition would probably be too vague or too limiting.

Over certain views there seemed to be no disagreement at all. No one disagreed that evaluation practice must ultimately extend beyond analysis of plausibility of proposed test score interpretation to include analysis of appropriateness of proposed test score use. Instead, the crux of the debate is whether the meaning of the word ‘validity’ should extend beyond plausibility to appropriateness; or, indeed, whether it should extend beyond truth to plausibility. As Borsboom and Wisjen put it, what is at stake is whether validity is best considered a matter of ontology, epistemology or ethics. Koretz, for example, insisted that validity is not a matter of ethics; if we wish to talk about broader evaluation issues, then we might refer to ‘unintended negative impacts’ (or suchlike), but the meaning of validity should be far narrower. Kane, on the other hand, suggested that validity cannot be separated from ethics; if we wish to talk about narrower evaluation issues, then we might refer to ‘evaluations of meaning-only interpretations’ (or suchlike), but the meaning of validity should be far broader.

There does seem to be some confusion concerning whether the Great Validity Debate is primarily lexical or conceptual. Newton and Shaw argued that it is

primarily lexical and that there is no logically correct answer to the question of how best to use the word. Markus also noted how the same validity theory can be expressed using different validity definitions, which expressed essentially the same point. Cizek, on the other hand, suggested that the debate is primarily conceptual and argued that the extension of validity to include ethical evaluation is simply logically incoherent. Kane implied that validity ought to be the 'bottom line in evaluating testing programmes' which, in turn, seems to imply that there is something powerful and compelling about the word itself.

The issue of who gets to call the shots in this debate arose both explicitly and implicitly. Borsboom and Wisjen constructed the argument as though between psychometric scientists (who tend to favour a narrower, more scientific view) and educational practitioners (who tend to favour a broader, more ethical view). They noted that educational testers do not own the concept of validity. Shepard seemed to imply that North Americans own the concept, when arguing that the (North American) consensus definition from the *Standards* should be defended on the basis that (North American) policy-makers, citizens and the courts have been taught it. Koretz, incidentally, observed that persistent attempts to teach a broad perspective on validity to stakeholders like these may not have led to understanding.

One potentially worrying observation from the papers within this special issue is the lack of consensus over the consensus definition. Geisinger actively promoted the *Standards* definition, arguing for a narrow concept of validity, and locating wider considerations within a broad concept of utility. Sireci also actively promoted the *Standards* definition, but claimed that the concept of utility was inherent within it. Whether these views reflected different conceptions of utility or different interpretations of the *Standards* definition is, admittedly, a little unclear, but the latter seems likely. Kane referenced the *Standards* when describing his 'consequences-as-indicators' model. This is a far narrower interpretation than the idea of testing the assumption that test score use does more good than harm, which is how Sireci characterised validation according to the *Standards*. Markus noted the importance of sensitivity to alternative vocabularies and to the possibility that authors might mean something other than what the reader might initially assume. Equally, though, there is surely an onus of responsibility upon authors to be as clear as possible what they mean by validity and validation, to minimise the risk of misinterpretation. Newton and Shaw noted that the *Standards* is quite ambiguous over the critical issue of how consequences relate to validity and validation. Zumbo and Hubley provided a detailed analysis which could be useful in helping future authors of the *Standards* to clarify the validity chapter.

Part of the contribution that a special issue like this one can make lies in helping to identify and clarify similarities and differences between alternative camps. The format of this particular special issue has allowed those classified within one camp or another to reflect upon those classifications and for authors to respond to those reflections. Both Cizek and Shepard cast some doubt upon the four-way classification of camps which Newton and Shaw proposed: ultra-conservative, conservative, traditionalist and liberal. Cizek took issue with the label 'conservative' and argued that the label (rather than the category, per se) best suited the position advocated by Shepard. The three-way classification of camps proposed by Kane was similar, although not directly overlapping: interpretation only model (similar to conservative); consequences-as-indicators model (similar to traditionalist); interpretation-and-use model (similar to liberal). Most controversial was the claim, from

Sireci, that a conservative-like position implied validation of 'useless' tests. Cizek insisted that this was simply a straw man argument. Real progress may be made if we are able to get to the bottom of radical differences of opinion, like this, concerning the detailed implications of different perspectives.

Of particular relevance in this respect is clarity concerning how far proponents of broader, more liberal perspectives are willing to extend their definitions of validity, and exactly why they are prepared to extend them thus far and no farther. Sireci, for example, promotes a fairly broad view of validity and validation, but presumably not as broad as that promoted by Moss. If not, then exactly how far does he go, and exactly why does he stop? Similarly, when Kane accepts that the range of social consequences may expand in the future, what kind of mechanism might provide the basis for this expansion, and according to what kind of criteria? Markus reminded us that there is an important difference between consensus based on non-deliberative majority rule and consensus based upon rational persuasion. By subjecting our agreements and disagreements to an iterative process of critical evaluation, clarification and refinement we will be able to move the field forward.

Finally, it is important to recall Gafni's observation that validation practice is often far from adequate and sometimes simply not conducted at all. We must not lose sight of the fact that there is far more to ensuring good validation than can be achieved by rigorous, scholarly debate over the meaning of validity.

Are we currently any closer to consensus than we were towards the end of the twentieth century? From a positivist perspective, we could comment that science moves slowly at times. From a postmodern perspective, we could argue that various interest groups and cultures have different positions. For example, the debate over whether consequences should be considered as part of validity often moves swiftly into the realm of who is responsible for those consequences. In highly litigious cultures, such as the US, this issue takes on a different complexion for the assessment industry than it does in other countries. In response to these issues, some of the authors have sought to use the logic of science to seek better definitions of validity or better arguments for particular claims relating to validity. There is a flavour of positivism sweeping through the special issue. Moss takes a distinctive approach in which she looks at the actual interpretations of test scores in use. In doing so, she shifts the focus of the debate from what validity is and should be (prescriptive theory) to what real-world phenomena validity theory needs to account for (explanatory theory). Given the state of the field, both of these approaches produce useful ways forward and the contrast between them helps to sharpen The Great Validity Debate.

Do we understand the issues at stake far better now than we did at the end of the twentieth century? We think that we do and believe that this special issue has made a significant contribution towards this end. The debate continues.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement, 50*, 110–114.

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

Crocker, L. (1997). Editorial: The great validity debate. *Educational Measurement: Issues and Practice, 16*, 4.
Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justification of test use. *Psychological Methods, 17*, 31–43.
Kane, M. T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.
Kane, M. T. (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement, 50*, 115–122.

Paul E. Newton
Jo-Anne Baird