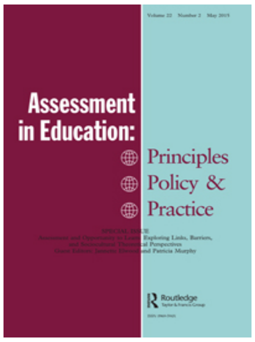

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

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

Validity, science and educational measurement

Harvey Goldstein

To cite this article: Harvey Goldstein (2015) Validity, science and educational measurement, Assessment in Education: Principles, Policy & Practice, 22:2, 193-201, DOI: 10.1080/0969594X.2015.1015402

To link to this article: http://dx.doi.org/10.1080/0969594X.2015.1015402

Published online: 01 Apr 2015.

Assessment in Education: Principles, Policy & Practice, 2015
Vol. 22, No. 2, 193–201, http://dx.doi.org/10.1080/0969594X.2015.1015402

Validity, science and educational measurement

Harvey Goldstein*

Graduate School of Education, University of Bristol, Bristol, UK

(*Received 7 August 2013; accepted 2 February 2015*)

The term 'validity' is one of the most important and one of the most debated concepts in educational measurement. In this paper, I argue that various different approaches can all be viewed from an associational perspective. I also argue that our understanding will be enhanced by adopting some basic ideas of scientific reasoning to the process of establishing validity and making fully transparent the assumptions and procedures used by both test constructors and test users.

Keywords: validity; associational validity; educational measurement; scientific reasoning

Introduction

One of the most important concepts in educational assessment, and also one of the most contested is that of 'validity'; the extent to which an assessment or a test instrument is considered fit for purpose. Historically, attempts to define validity have undergone many changes (see Lissitz, 2009 for a review), and while a debate continues, at present a consensus, at least among the psychometric community, appears to exist on the nature of validity. According to Lissitz (2009, p. 20), this consensus regards validity as concerned with the interpretation of a test score[1] and not as an inherent property of the test itself. He considers that a user is seeking to validate one or more interpretations made from a test, not the test itself. The North American standards for educational and psychological testing (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1999, p. 9) concur with this.

To the casual user, such a view about a central component of a field that is concerned with how to measure almost anything that one might want to make a judgement about, may seem curious. After all, if we make claims to be able to measure some very subtle mental processes, why do we seem to have given up on *measuring* the validity of those same instruments? In fact, it has not always been so. Historically, there have been attempts to produce 'validity coefficients' for tests, based upon association measures such as correlations with other tests or judgements made concurrently or in the future of an individual's life. I will refer to these as 'associational validity' and they include 'concurrent validity', 'predictive validity', etc. The striking thing about the current consensus definition is that psychometrics appears to have given up on attempts to provide quantitative measures of validity. Instead, debate revolves around the nature of the 'constructs' (including content)

*Email: h.goldstein@bristol.ac.uk

within a test and how these are ‘consistent’ with any use of the results. Thus, according to this consensus, any given user of a test is asked to make a judgement, implicitly if not explicitly, using all available relevant evidence. In other words, validity becomes a value judgement that takes in all aspects of the testing situation. As such it is legitimate that it may differ among users who may have different purposes and views: what is important is that these are made transparent.

In this paper, I want to look at the implications of this view and to suggest that if it is to be taken seriously, test constructors and especially those who wish to claim a firm, ‘theoretical’, quantitative basis for testing need to think rather carefully about what this notion of validity implies for their craft. In particular, I shall develop an argument that views validity as a process that needs to conform to accepted scientific norms of falsifiability and replicability.

Validity by association

Let me start with a ‘simple’ example to illustrate where I am coming from. Once a test is in existence, say an arithmetic test for 10-year-olds, a user of the results will want to have some reassurance about its ‘validity’. They may decide that it is not suitable for their purpose in which case they may wish to modify it, choose another or simply not bother with testing arithmetic at all and rely upon some other type of judgement. How might they do this?

Let us suppose that the test is used in order to decide how to ‘stream’ children into different learning channels and also suppose that results are available for a group of children in one school. The responsible teacher might decide on several kinds of action. To begin with she may decide simply to accept the test scores at face value, especially if it comes from a ‘reputable’ source such as a large test publisher or if it is one mandated by government advisors. If we go along with the consensus then the use of the test for streaming will be judged as valid (by that user). Of course, a different teacher may decide differently, but there are no guidelines for deciding one judgement is more worthy than another. The same is true if the judgement is based upon the (possibly conflicting) views of several teachers – there is always the possibility that people may view things differently. In other words, unless a rule is imposed that requires conformity to a particular judgement, for example that of the most ‘senior’ assessor, or according to existing policy, an element of subjectivity has to be accepted. Note also that the *context* of these varying judgements is the same; it is the views of the judges that differ. And of course they may differ for all sorts of reasons – the APA standards and other commentators are not prepared to rule out any particular type of judgement; all that is required is a reasoned ‘argument’ to support the judgement. Certainly, there seems to be no attempt to define the person or kind of person who is or is not allowed to pass judgement nor is there any attempt to rule out certain kinds of judgements such as those based upon ‘social’ consequences. Thus, it would be legitimate, according to the consensus view, for a test that exaggerated the scores of White children against Black children to be described as valid if that was what a particular teacher wished and if they could support it with a well-grounded argument. Many might argue that such a user should not be allowed to practise, and indeed there might be a law against it, but the test itself would be valid for that particular person and that particular purpose. Newton (2012) makes this explicit when he points out that validity should be judged by the *stated* interpretation that is placed upon a test by those who publish and also by those who use it, and this transparency does seem to be crucial.

But note what is being said in this example. Validity is being judged according to a particular *outcome*. In other words, this is a (crude) form of associational validity – the test is associated with ethnicity. In fact, my point is a more general one, namely that the consensus definition of validity cannot escape the confines of a definition that ultimately rests upon measures of associational validity – even if these are crude and implicit. In the case of a test for streaming, the fitness or validity of the test will be determined in terms of how well it is associated with optimum placement of students in streams – if it is seen to fail to do this well then it would have low validity. Of course, the 'criterion' has to be chosen, such as association with ethnicity or streaming, but that is the case with any of the historical measures of associational validity.

So now let me look at some of the contemporary debates about validity in the light of my claim. Following that review I will conclude by arguing that we can retain judgemental aspects of validity, while at the same time seeking to 'objectify' them as far as possible by having a transparent framework for communicating the assumptions made by test constructors and how these come to be interpreted by users when judgements are made.

Current debates about validity

Implicit in the APA standards, and also in the influential writings of Messick (1989), is the centrality of the notion of 'construct' validity. This refers to the meaning of the test in the light of how it is used, and requires an understanding of how the test was constructed and how testees interpret what they see. Yet this also, I would argue, is predicated upon appropriate associations. The 'meaning' of a test has to have some empirical content. Our arithmetic test consists of items whose meanings will depend on how responses to them actually (as opposed to theoretically) occur. If a child who is *perceived* as being arithmetically competent performs poorly on an item this suggests that the item needs modification and in fact the process of test construction involves just such kinds of iterations between proposing appropriate items or collections of items and a judgement, formal or informal, about whether they correspond to existing notions of what they should be reflecting. I will return to this issue of 'existing notions' later. In other words, simply talking about 'construct validity' cannot hide the actual dependence of test construction and interpretation on some form of associational validity. Indeed, short of generating test items in some arbitrary or random fashion, it is very difficult to see how any test could be constructed that was not meant to reflect some association with pre-existing concepts. This is not to say that there is no theoretical component to test design or to understanding test relevance in any given context. Choice of items, their format, wording, etc. may well be inspired by educational or psychological theory, but the *validation* of the test has to make an appeal to associations.

Before I move on to considering a general framework, let me look at some of the existing debates in the light of the above.

First of all, the need to appeal to associations is echoed by many commentators. Thus, Sireci (2009, p. 32) states that tests should 'demonstrate predicted relationships with other measures of the intended constructs'. Zumbo (2009, p. 69) suggests that 'validity is the explanation for the test score variation'. I take him to mean that differences between individuals or groups of individuals are associated with other characteristics that are relevant to what the test is meant to measure. In practice, therefore, this implies an associational validation process. Mattern, Kobrin, Patterson, Shaw, and Camara (2009, p. 216) are very clear that validity for the US College Board Scholastic Aptitude Test resides to a large extent in its ability to predict later performance. Mislevy (2009, p. 104) also emphasises the importance of 'correlations with other data and predictions of criterion performance'.

Borsboom (2005) stands out as claiming to take a rather different approach to validity from the majority of commentators. In his view, validity is a property of a test and a test is valid if variations in test scores are directly related to the attribute that is being tested. It relies upon the ontological assertion that there really is an attribute that is causally related to test performance. He explicitly denies that association has a role to play and states that, 'criterion validity was truly one of the most serious mistakes ever made in the history of psychological measurement' (p. 159). His claim is that the process of measurement is that of the measuring instrument varying as a result of some causal process operating through an individual. It is difficult to understand his argument, however, since he clearly believes that measured associations can be modelled in order to infer causal relationships. Thus, although Borsboom appears to take a different position to other commentators, in order to operationalize his views, he still relies upon observed empirical associations, since the absence of any such associations (after allowing for possible confounders) would certainly not allow causal inferences.

It is quite instructive to take Borsboom's example of a test of Piaget's theory of developmental stages for children. He describes a task (p. 165) involving balance weights and describes how responses can be modelled to validate the theory. What he omits to say is that this validation relies upon the observation that children actually do pass through these stages in a common order. In other words, the observation of the stages is correlated with age.

Finally, the work of Cronbach, and especially the paper by Cronbach and Meehl (1955), has been especially influential. Their support for the notion of construct validity has influenced many subsequent developments. Thus, they suggest that construct validity is involved 'whenever a test is to be interpreted as a measure of some attribute or quality which is not "operationally defined"' (p. 282). In practice, they are clear that associations play a key role and they implicitly advocate a form of hypothesis falsification by speaking about rejecting competing explanations for observed associations, and I shall return to this below. The main part of their paper is concerned with how observed patterns of association can be interpreted to infer validity. They conclude: 'A construct is defined implicitly by a network of associations or propositions in which it occurs' (pp. 299–300).

Implications and science

In the preceding section, I have argued that the differing views of what constitutes test measurement validity all rely upon the existence and interpretation of associations between the test score and other criteria that may be relevant to the underlying 'construct' or the use to which the test is put. There are, of course, differences in the various approaches to validity and these may be important for the ways in which tests are constructed and used. Our test of arithmetic might be constructed by attempting to sample at random arithmetic questions from a 'population' of such items, or it may be derived by selecting those items from a candidate set according to those which most strongly discriminate between a group of children divided by teachers into high and low arithmetic performers, or the items might be selected on the basis of which items a test developer considered best reflected the construct 'arithmetic'. When it comes to *validating* the test, however, the only way that this can occur is through an understanding of how actual test scores relate to 'arithmetic' as it is exhibited, and modelled through explicit or implicit comparisons with external criteria. The implications of this will now be explored in more detail.

The first point to note is that there is no important distinction between what might be termed implicit and explicit measurement of association. An association may be observed informally or by appealing to claims made for the way in which test items have been found to relate to performance on tasks or derived from observing examples of how 'arithmetic' is taught. Informal procedures may not be felt to be very robust or replicable, but they are still based on the idea of association with external criteria. Traditional notions of 'predictive' or 'concurrent' validity of course are direct attempts to measure association with an external criterion, but their weakness has been perceived as being too narrow. Typically, such criteria have often been too limited in scope, difficult to generalise across contexts, and too vague about what constitutes acceptable levels of association. Thus, Cronbach and Meehls' proposal in favour of 'construct validity' is to extend the associations into what they describe as a 'nomological net' by which they mean a set of quantitative relationships and a 'chain of inference' that allows validity to be established. Mislevy (2009) echoes this in his discussion of latent variable models that seek to explain observed relationships. I do not wish to argue particularly in favour of or against such model-based attempts per se. What I do want to do is to set the debate within a more general scientific framework and to explore the consequences of so doing. If it is accepted that to talk about validity is essentially to discuss, model and make inferences based upon explicit or implicit associations, then the nature of these associations has implications for those inferences.

There are two key features of a scientific approach within any discipline. The first is to do with falsifiability, as already mentioned, and the second with replicability. The falsifiability criterion (Popper, 2002), briefly, is that a scientific process should be seeking, continually, for evidence to discredit or falsify whatever hypothesis is currently used to explain a phenomenon. Elsewhere (Goldstein, 2012) I have discussed this with respect to much of modern psychometrics that elevates verification above falsification in terms of its practices. The use of 'goodness of fit' tests for item response models and 'confirmatory factor analysis' are examples of this where falsification typically is neither advocated nor attempted. Instead, attempts are made to show how quantitative results are *consistent with* the hypothesis. So, in terms of associational validity what would falsification consist of?

First, it would suggest that if we wish to validate a test of arithmetic then we should actively seek associations with our test scores that appear to run counter to what we would expect. Thus, the ranking produced by the test may not coincide with that suggested by a teacher. This may remain the case over replications of the test with many different teachers and children. If so, this may be taken as evidence that the test was inadequate and should be modified. A process of modification and testing could then be embarked on until a more satisfactory agreement was established. At this point, the process of falsification would not cease, although it might be set aside pro tem for practical reasons. A new stage of validation might involve using the test in very different contexts to see how well its associational properties were sustained. Such replication might suggest a widespread applicability or perhaps a very context-specific validity. New criteria might be sought against which it could be assessed, such as other tests also claiming to be tests of arithmetic for 10-year-olds. In other words, the validity of a test will have the same status as any other scientific hypothesis. Nevertheless, there remains an important issue which is that the criteria for judging validity may be contested so that different users might have different claims for validity based upon the same evidence. This of course can happen in other sciences, but in more mature disciplines agreement will eventually be achieved. This need not happen, and indeed often does not happen, in the social sciences and education, itself a somewhat immature discipline, in particular. Different world views and ideologies about what should count as knowledge are of the essence.

An enlightening illustration of this occurred in the 'Golden Rule' case (Goldstein, 1989). The Golden Rule insurance company noted that the test it was using for insurance recruitment, supplied by Educational Testing Services (ETS), seemed to be rejecting too many Black applicants. After some negotiation, ETS agreed to supply a test where items were selected, at least in part, on the grounds that they minimised Black–White differences. The subsequent history is that ETS broke the agreement, and argued that such a consideration was not appropriate for constructing a test. The point, however, is that the insurance company was simply exercising its view about associational validity – that for its own purpose Black–White differences should be minimised. We do not know what implicit or explicit associations were used by ETS to construct its original test, but there would have to have been some. One of the arguments used by ETS was that political or social considerations should play no part in test design, but we do not know how their original test was influenced, consciously or unconsciously by such considerations, especially in terms of the test constructors' views about expected or acceptable Black–White differences. In an ideal world, however, these assumptions would be made explicit and available for inspection so that any particular user could decide to accept or attempt to modify them, and organisations such as the Golden Rule insurance company could argue their own case for establishing validity.

In some ways what I am proposing is not too far removed from the view that the notion of validity has to be contextualised for every user and is not of itself an inherent quality of the test itself. I am, however, going further than that. I am suggesting that, if we must use the term, then validity could indeed be regarded as a characteristic of a test, or not as the case may be, but that this needs to be subject to a process of deliberate falsification with a variety of alternatives and to a process of replicability. The historical debates about whether validity is a 'process' or a test characteristic are interesting but not fruitful. If we wish to think of a test as measuring *something*, whether this is labelled a 'construct', a trait or whatever, then what we need to do is provide empirical evidence for the things it is associated with and search for those things that are capable of undermining our interpretations.

Finally, if we wish to make any kind of general claim for validity, our test instrument should be replicable within the contexts for which it has been designed. This implies not simply that we attempt to falsify it in different contexts but also that the results of using it, its consequences, are similar.

Things as they are

Let me end with some remarks on how test construction appears to operate. It has been pointed out that the users of tests are by and large separate from the constructors of tests. Thus, in England, national curriculum tests are devised by small groups working to a strict timetable and set of protocols. The users of the tests are teachers in schools and the consequences that flow from the results of the tests have important effects on those teachers and schools – typically in the form of league tables and targets. In the USA, the 'No Child Left Behind' legislation performs a similar function.

One of the things we know about the construction of many such tests is that they are designed to be as comparable as possible across time. In particular, expectations about group differences will be incorporated into them. For example, when formats change, different groups may be advantaged such as the case where girls seem to have an advantage when test items are free format rather than multiple choices. If this occurs, it is often viewed as unfortunate and a threat to the validity of the new instrument. In other cases, procedures that would have the effect of altering more subtle relationships may also be discouraged. What I am suggesting is that built into the test construction process, there is often a rather conservative element that attempts to ensure that any new test produces similar results to previous tests, whether this is explicitly engineered or appears via an implicit process. We often think we 'know' about certain group differences. As Gould (1981) pointed out, however, historically such knowledge can be determined in subtle ways by unquestioned cultural assumptions. He has an interesting example of head measurers in the late nineteenth century who consistently found that the rank order of head sizes in different groups followed exactly pre-existing 'racial' assumptions so that White males had the largest and Black females the smallest. What Gould was able to show was that these assumptions led them to judge the 'validity' of potentially suspect measurements in ways that reflected their assumptions and biased their results. When the data were re-analysed more neutrally, the racial differences disappeared.

More contemporaneously one might ask, for example, whether the often taken-for-granted differences in 'spatial ability' between males and females could have arisen, at least partly, because of assumptions built into the very first such tests that expected to find males scoring higher than females. The choice of tasks may affect differences due to differential exposure that may be linked to specific cultural contexts. Thus, for example, males are often found to outperform females in spatial tasks involved with video games, but Feng, Spence, and Pratt (2007) found these differences reduced following a period of familiarity with playing such games. Likewise, Murphy (1991) elaborates on the different types of solutions boys and girls bring to problem-solving tasks; in particular, boys' relative unwillingness to abandon an ostensible 'real-life' context in favour of an abstracted technical issue. The issue for test constructors is to decide just what is relevant, and in particular the test constructor cannot escape from making either an explicit or implicit judgement about whether to advantage, say, boys or girls. Simply to assert 'technical' compliance with a set of standards is to avoid the question. Indeed, as I have suggested in the Golden Rule case, the test constructors may have allowed implicit judgements about expected racial differences to influence test content even though the testing agency was not prepared to admit the relevance of this.

To claim validity, a test instrument needs to imply testable, falsifiable, consequences that can, in principle, be evaluated empirically. I have argued that these will consist of associational relationships to other measures or judgements. I would propose, therefore, that any test that seeks to be regarded as having a degree of validity should be required to state clearly what such relationships are considered to be relevant so that its claim can be evaluated. Thus, our arithmetic test needs to state what it is expected to be associated with. As part of this, the assumptions that formed part of its construction process need to be set out. For example, was it designed to parallel an existing test? How were items accepted and rejected – not simply on grounds of 'fit' but in terms of how they were perceived to be associated with other measures? Were items that failed to exhibit 'expected' gender differences or ethnic group differences as in the Golden Rule case, rejected? Many assumptions have to be made during the test construction process and requiring constructors to be explicit about these will aid the process of test validation. I would suggest that they are incorporated into test standards. Above all I am suggesting a shift of emphasis away from what are often somewhat obscure philosophical debates about the nature of validity, towards a recognition that common standards of scientific judgement are of major importance and should be applied routinely.

Acknowledgements

I am most grateful to Jannette Elwood, Patricia Murphy and Paul Newton for helpful comments.

Note

1. Throughout this paper, I shall use the term 'test' to refer to an assessment instrument that is reusable by different assessors, rather than the broader meaning of any procedure or device to make a judgement about individual competence or learning, such as an informal teacher rating.

Notes on contributor

Harvey Goldstein is a professor of Social Statistics at the University of Bristol. His research interests include quantitative models for assessment, multilevel modelling and methodological tools for social science data analysis. His major recent book *Multilevel Statistical Models* (Wiley, 2011) is a standard reference work.

References

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

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

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

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science, 18*, 850–855.

Goldstein, H. (1989, March 27–31). *Equity in testing after Golden Rule*. Paper read to the American Educational Research Association annual meeting, San Francisco, CA.

Goldstein, H. (2012). Francis Galton, measurement, psychometrics and social progress. *Assessment in Education: Principles, Policy & Practice, 19*, 147–158.

Gould, S. J. (1981). *The mismeasure of man*. New York, NY: W. W. Norton.

Lissitz, W. L. (Ed.). (2009). *The concept of validity.* Charlotte, NC: Information Age.

Mattern, K., Kobrin, J., Patterson, B., Shaw, E., & Camara, W. (2009). Validity is in the eye of the beholder: Conveying SAT research to the general public. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 213–240). Charlotte, NC: Information Age.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). Washington, DC: American Council on Education.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 83–108). Charlotte, NC: Information Age.

Murphy, P. (1991). Assessment and gender. *Cambridge Journal of Education, 21*, 203–213.

Newton, P. (2012). *Clarifying the consensus definition of validity.* Cambridge, UK: Cambridge Assessment.

Popper, K. P. (2002). *Conjectures and refutations. The growth of scientific knowledge*. New York, NY: Routledge.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Charlotte, NC: Information Age.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 65–82). Charlotte, NC: Information Age.