
Peer reviewed version

Link to published version (if available): 10.1080/0969594X.2015.1015402

Link to publication record in Explore Bristol Research
PDF-document

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/about/ebr-terms
Science, validity and educational measurement.

Harvey Goldstein, University of Bristol

h.goldstein@bristol.ac.uk

Abstract

The paper suggests that our understanding of educational assessment is enhanced by adopting some basic ideas of scientific reasoning, especially to the process of establishing validity and making fully transparent the assumptions and procedures used by both test constructors and test users. The concept of validity is explored in depth as is the contexts in which it may be expressed. It is argued that the various different approaches to defining validity can all be viewed from an ‘associational’ perspective.

Keywords

Associational validity, educational measurement, falsification, scientific reasoning, validity,
Introduction

Those involved in educational assessment as an intellectual activity would typically wish it to be viewed as conforming to established scientific principles. This is perhaps most obvious in the area of test construction where formal guidelines have been developed (see for example, APA, 1999) and where formal mathematical and statistical modelling underpins mainstream activity. It also applies, however, in many less formal assessment environments such as formative assessment where there are concerns to establish reliability, replicability and validity in order for inferences to be acceptable. In this paper I will explore how scientific reasoning can inform an understanding of educational assessment and in particular how different approaches to the concept of ‘validity’ can be viewed from such a perspective.

There are two key features of a scientific approach within any discipline. The first is to do with falsifiability, 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. For example, the use of ‘goodness of fit’ tests for item response models and ‘confirmatory factor analysis’ are examples where falsification typically is neither advocated nor attempted. Instead, attempts are made to show how quantitative results are consistent with an existing hypothesis.

In terms of replicability, assessment shares a particularly difficult problem with many areas of the social sciences, namely that the context is an important determinant of measurement outcomes. An assessment instrument may have quite different properties depending on who it is given to, who administers it, where and when it is administered and its perceived purposes. All of this is well known but often ignored. Thus, for example, formal examinations originally developed for certification purposes may come to be used for accountability purposes as in school performance or ‘league’ tables, and the knowledge that this happens will change the way students and institutions respond to them (Foley and Goldstein, 2012). Likewise, different groups such as males and females may exhibit different responses according to the precise format of test questions or the real life contexts in which they are embedded. Such factors typically are ignored by test constructors who generally like to claim that the instruments they produce have a common currency over a wide range of contexts and the term ‘standardised test’ reflects this assumption.

The concept of ‘validity’ is central to all assessment and has been a topic of debate from the early years of the 20th Century. In this paper I will look at this concept in detail and judge the various claims about validity against the scientific requirements for falsifiability and replicability.

Validity

Broadly speaking, validity is the extent to which an assessment or 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, P20) this consensus regards validity as concerned with the interpretation of a test score and is not an inherent property of the test itself. He considers that a user should seek to validate one or more interpretations made from a test rather than to validate the test itself. The North American standards for educational and psychological testing (APA, 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 the measuring instrument. 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) 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 even though it can differ among users who may have different purposes and views: what is considered to be important is that these are made transparent.

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 reassure themselves?

Let’s 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 view 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 that one judgement is more worthy than another. People may view things differently. In other words, unless a rule is imposed that requires conformity to a particular judgement, for example that of a senior assessor, or according to existing policy, an element of subjectivity has to be accepted. Note also that the context of these varying judgements may be the same, it is the views

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.
of the judges that differ. 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 could support it with a well-grounded argument. Many might argue that such a user should not be allowed to practice, 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.

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 – for example that 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 associational validity. Following that I will argue 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, and all of this subject to the process of falsification and replicability.

**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 would generally suggest 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. This is still the case where formal modelling, for example item response modelling, is used as part of the test
construction process, where the test constructors’ assumptions are translated into mathematical assumptions that determine the form and content of the model. Moreover, the theory, assumptions and knowledge that test constructors use are themselves based upon empirical observation.

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 purely 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 can be no theoretical component to test design or to understanding test relevance in any given context. Thus, for example, if the validity criterion is a (subjective) perception of mathematical competence, there is no necessary requirement for this perception to be theoretically grounded, although this is not ruled out. Such things as the choice of items, the number of underlying dimensions, 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. Furthermore, there may or may not be a causal link involved, but the existence of such a link is essentially a separate issue that will generally involve at least some empirical study, as pointed out below in the discussion of Borsboom’s views.

The need to appeal to associations is echoed by many commentators. Thus Sireci (2009, P32) states that tests should ‘demonstrate predicted relationships with other measures of the intended constructs’. Zumbo (2009, P69) suggests that ‘validity is the explanation for the test score variation’. I take him to mean here 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 et al (2009, P216) are very clear that validity for the US College Board Scholastic Aptitude Test (SAT) 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 than 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 (P159) ‘criterion validity was truly one of the most serious mistakes ever made in the history of psychological measurement’. 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 for causal inferences. In fact, 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 (P165) 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 that the observation of the stages is associated with age.

Finally, the work of Cronbach, and especially the paper with Meehl (Cronbach and Meehl, 1955), has been especially influential and is strongly supportive of an associational perspective. 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 which is not ‘operationally defined’. 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. 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’.

Validity and individuality

Almost all theorising about assessment is concerned with describing the performance or competence of individuals. Yet individual learning typically is a social experience carried out through interactions with others – both teachers and learners. Likewise, the outcomes of learning are often at the social or group level. Yet we hardly ever assess at such a level, let alone relate the social and individual outcomes of learning. This failure to account for an important context of learning has implications for validity.

Traditional models of assessment are thus essentially individualistic. The traditional psychometric and educational assessment paradigm is that of an individual confronted by a set of questions or test items to which they are expected to provide responses. The responses that are obtained are then treated as indicators of some attribute, or attributes, residing within the individual, either permanent as in notions of innate intelligence, or contingent as in most kinds of educational attainment measures. This is the case whether assessments are ‘summative’ or ‘formative’. For example in the latter, the emphasis is on assessing what each individual understands or can perform, even though such approaches tend to pay more attention to context such as peer group influences. Thus, for example, in the work of Black and Wiliam (1998) the evaluation of their claims for learning enhancement rely on ‘effect sizes’ based upon testing and examining individuals. Likewise, in his recent book, when Stobart (2008, P 145-146) describes ‘Assessment for Learning’, he appears to be concerned essentially with the learning patterns of individuals. In all these cases the key assumption is that, when it comes to making assessments, each individual is treated independently, as if their responses are unaffected by the responses of other individuals or contingent aspects of their environment.

Clearly, the assessment of individual understanding and performance is important for a whole number of purposes. At the same time, however, the functioning of groups is, in many situations, just as important if not more so. Thus, teamwork may be highly valued and the performance of the team may be paramount. In many sporting activities the team is the primary focus of assessment, even when the individual members may change. In classrooms, group work may result in assessments being made at both group
and individual level rather than just at individual level, and there are many other examples.

When a set of individuals is assessed, it is common to summarise the assessments in terms of one or more aggregated results. Thus, for example, an average rating may be made for a group studied and assessed by a teacher over a period of time, or average examination results may be computed for a school. Comparisons of teachers or schools might then be carried out using such averages. In the case of summative assessments this is often the primary purpose but, even where assessments are formative, there may be an interest in making comparisons for different groups of students, or the same students across time.

A standard feature of such comparisons is the assumption that the rating or score achieved by any particular individual has no effect on those obtained by other individuals, other than through shared exposure to, for example, teaching or contextual factors. This is the traditional assumption of independence already mentioned. At the level of schools, likewise, comparisons among schools assume that the schools produce results independently, after making any adjustments for factors such as intake and shared context. This is a key assumption, for example in value added analysis of school test score results (Leckie and Goldstein, 2009). Nevertheless, such assumptions can be questioned, and with this claims made for validity. Consider first the case of schools.

Assume, for the sake of argument, that school examination scores are affected not only by the characteristics of their students and teachers but also by the resources they receive, including real monetary ones and those non-monetary ones such as provided by parents; some evidence for this exists (see Steele and Vignoles, 2006). Where schools are competing for resources, for example in cases where resources are allocated partly on the basis of student test performance, an uneven distribution will generally result. Suppose also that the total amount of resources is fixed and also that we have no good measure of how many resources of all kinds are actually received by schools. We then have a situation where an increase in resources, associated with a change in outcomes, for one school will be related to a change in outcome for the remainder since these will suffer a drop in resources. This will therefore lead to a lack of independence among the outcomes. The consequences of ignoring this dependency will depend on the nature of the dependency, but will generally have a biasing effect on estimates of variation between students and between schools and also lead to underestimation of standard errors and thus biased significance tests and confidence intervals. This is analogous to the well known effect of fitting a single level model when in reality a multilevel structure is present (Goldstein, 2011).

Now consider the case of individual students. We have, in principle, the same set of problems. Thus, for example, if we think of teacher time as a resource, and one whose allocation in general is difficult to measure accurately, then for similar reasons this will induce dependency among student outcomes. Another issue occurs when groups of students collaborate, resulting in dependencies, but for a somewhat different reason. In
this case, the existence of collaborative groups is rather like the existence of separate schools into which children are grouped so that performance within a school is generally more similar than the performance of students in different schools. A failure to take account of this will result in underestimates of standard errors and confidence intervals and this will generally be handled by fitting a multilevel model that recognises the grouping dependency. In principle we could also fit a multilevel model to individual student data where grouping occurs, but the difficulty will lie in knowing how the groups are constructed, and how they may reform over time.

What I am reiterating here is that the validity of an assessment needs to take account of the context within which it is made. To ignore such contexts, or to assume that they do not matter, may distort what we are trying to measure. An individual responding to an assessment in one context or one peer group may give quite a different response in another. When a validity statement is attempted it needs to describe the context within which the assessment occurs and a task for the assessment community is to provide tools for making such descriptions.

Scientific implications

I have argued that the differing views of what constitutes test measurement validity all rely upon the existence and interpretation of associations between the assessment outcome 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 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 predefined ‘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 theoretically important distinction between what might be termed implicit and explicit measurements 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 traditionally their weakness has been that they are perceived as being too narrow. Typically, such criteria have often been too limited in scope, difficult to generalise across contexts and so with low replicability, and too vague about what constitutes acceptable levels of association.

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 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.

First, in the spirit of falsification, 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 might reasonably 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 – potentially falsifiable.

Yet, there remains an important issue which is that the criteria for judging validity may themselves 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 is more likely to be achieved. This need not happen, and indeed often does not happen, in education, itself a somewhat immature discipline. Different world views and ideologies about what should count as knowledge are of the essence.

An enlightening illustration of opposing ideological assumptions 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 and what counted as a criterion for establishing 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. The testing industry may well find it uncomfortable to acknowledge that there is a subjective, value-based, element in the construction of assessment instruments, from those designed to evaluate specific learning programmes through to ‘general ability’ tests. Nevertheless, if the different (subjective) assumptions were made explicit, as happened in the Golden Rule case, this would, in my view, enhance the assessment enterprise.

An interesting example of the importance of a falsification perspective arises in the dimensionality assumptions made when using statistical item response models to construct tests. The simplest such model which has achieved a wide currency for both test construction and analysis is the so-called ‘Rasch’ model. In its most basic form it assumes that a set of ‘binary’ items that allow correct or incorrect responses are available, and that all or some of them are to be used to form a test, for example of mathematics achievement. The Rasch model is fitted to a sample of individual responses to the items and the way in which the items relate to the fitted model determines which ones will be used in constructing the test. One of the key assumptions built into the model is that the (correct/incorrect) responses are determined by a single underlying dimension, and items that do not conform are discarded or modified until they ‘line up’ with a single dimension. More complex versions of this model also make a similar assumption. Proponents of such modelling procedures will often use weak ‘goodness of fit’ tests to demonstrate that their model ‘fits’ their data rather than attempting to falsify directly the assumption of unidimensionality by fitting more complex models that incorporate two or more dimensions. Goldstein (1980) shows how a mathematics test fitted using a Rasch model is better understood by incorporating two dimensions that reflect different aspects of mathematics. Goldstein et al. (2007) show how reading performance scales used in the international PISA study of reading performance, based upon the Rasch model, are in fact at least two-dimensional, and that fitting such higher-dimensional models provides greater insight into the data. Both these examples illustrate the how the interpretation and hence the validity of a test, is effectively undermined when adopting a ‘confirmatory’ rather than a falsifiability approach.

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 also to a process of defining the contexts within which it is replicable. As with other scientific endeavours, the more a test survives attempts at falsification of its claims the more acceptable it becomes. 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.
Where are we now?

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 United States 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 and this can lead to a kind of ‘historical determinism’. 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 choice. If this occurs it is often viewed as unfortunate and a threat to the validity of the new instrument. 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 (1982) 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 19C 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 the measurers 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 contemopranently 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 out-perform females in spatial tasks involved with video games, but Feng et al. (2007) found these 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 willingness to abandon an ostensible ‘real-life’ context in favour of an abstracted technical consideration, and Elwood (2005) discusses how exam format and content affect differential gender performance. Elwood (2006) discusses the way in which expectations about gender differences affect teacher judgements. She also makes the point that these expectations will affect the learning process as well as assessment results. This in turn will affect validity judgements as I have argued earlier and also supports the idea of an historical determinism.

The issue for test constructors is to decide just what is relevant, and in particular the test constructor cannot escape 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 key 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.

I have argued that to claim validity a test instrument needs to imply testable, falsifiable, consequences that can, in principle, be evaluated empirically. Test constructors and/or users also should be able to specify the contexts over which its results can be considered replicable. Furthermore, any test that seeks to be regarded as having a degree of validity, should be required to state clearly what associational relationships are considered to be relevant so that its claims to validity can be evaluated. Thus, our arithmetic test needs to state what it is expected to be associated with. As a component of this the assumptions that formed part of its construction process will 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? What assumptions about dimensionality were made? 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 these requirements are incorporated into test standards. Above all I am suggesting a shift of emphasis away from what are often somewhat obscure debates about the nature of validity, towards a recognition that openness about assumptions and adherence to common standards of scientific judgement are of greater importance and should be available routinely.

Acknowledgements

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


