

**THE VALIDITY OF ABILITY TESTS – A CASE OF
OVER-INTERPRETATION?**

GORDON STOBART

INSTITUTE OF EDUCATION UNIVERSITY OF LONDON

email: g.stobart.ioe.ac.uk

The validity of ability tests – a case of over-interpretation?

The title of this paper is not self-explanatory and so needs some immediate decoding. In current theorising about validity, which I describe later, the focus of a validity inquiry is on *the inferences made from the results*. So it is not the test that is valid, it is what you do with it that determines validity. This includes its impact, particularly in relation to the purpose of the test.

A validity argument is about how well a test serves the purpose for which it is used. If there are multiple purposes each needs to be evaluated (AERA, 1999 Standard 1.1); a test may be more valid for one purpose than another.

So it is what is inferred from ability tests that is of interest here, though for those with no interest in them there is a general issue about the naming of any test. My basic argument is that too much is inferred from scores on ability tests – they are *over-interpreted*. This is because the term ability carries ‘excess meaning’ through its association with notions of fixed intelligence and IQ tests. These extra layers of interpretation reduce the validity of ability tests because they cannot support such generalisation. I will develop this argument later, for the moment an analogy might help. If I interpret the results of a 20 metre swimming test to claim someone can swim 20 metres, the validity argument becomes relatively straightforward. If I go on to use the same results to identify potential Olympic swimmers I have a much bigger validity claim on my hands. This would be a case of over-interpretation: reading more into the results than they can support.

I also need to make clearer what I mean by ‘ability testing’. This is not straightforward because definitions often incorporate the excess meaning I am seeking to resist. Like the 20 metre swimming test there are two levels of interpretation of what an ability test measures:

1. generalised achievement associated with schooling. It is not directly curriculum related (unlike achievement tests such as examinations) but draws on broad skills and knowledge (vocabulary, numeracy, reasoning);
2. general intellectual capacity which is independent of schooling and which determines achievement. On this basis ability is the cause of learning rather than the outcome.

My argument is that the validity claims of 1 are much stronger than those of 2, in which over-interpretation has taken place. My concern is that, in teachers’ discourse in England, it is the second use of ability that looms large. Ability is something you enter school with (or without) rather than gain through schooling.

To develop my approach in more detail I begin by looking at current understandings of validity, which I then apply to ability testing. I use as my examples of the Cognitive Ability Test (CAT3) in England, a test taken by two-thirds of 11 year olds on entry into secondary school, and the college admission SAT test in the US.

Validity

There are some clear international differences in the importance attached to the construct of validity. In the American *Standards for educational and psychological testing* (1999) validity assumes pride of place as ‘the most fundamental consideration in developing and evaluating tests’ (p 9). It is the first section in the *Standards* with 15 pages of devoted to the 24 validity standards. This reflects a long tradition of theorising validity in testing which can be traced for over 50 years through the different editions of the *Standards* and through successive chapters in the four editions of *Educational Measurement*, the latest being Kane’s in 2006.

In the UK, unlike the United States, validity has never been a major theoretical debate. In his classic 1991 analysis of the English examination system Robert Wood commented:

The examining boards have been lucky not to have been engaged in validity argument. Unlike reliability, validity does not lend itself to sensational reporting. Nevertheless, the extent of the boards' neglect of validity is plain to see once attention is focused. Whenever boards make claims that they are measuring the ability to make clear reasoned judgements or the ability to form conclusions, they have a responsibility to at least attempt a validation of the measures...Validation work is unglamorous and needs to be painstaking but has to be done. As long as examination boards make claims that they are assessing this or that ability or skill, they are vulnerable to challenge from disgruntled individuals (pp. 151-2)

In the 2007 QCA Code of Practice for national curriculum assessment in England, validity is mentioned 17 times in relation to a variety of concerns (for example fairness, malpractice, script scrutiny, level setting) rather than being a central organising principle¹. Validity can be seen in the Code of Practice as a largely implicit and under-defined concept - 'the fitness-for-purpose of an assessment tool or scheme' (p 75).

I will therefore align with the American tradition, which has in part been driven by the need to show legally defensible processes of test construction. This historical process has, in essence, been one of ever widening interpretations of validity and the logic of this evolution is worth considering. Validity has evolved from being a statistical property of a test, to articulating the construct and content involved, which then led to making the use and consequences of the test *results* central. In this evolution the key transformation has been the move away from treating validity *as measurement of a fixed property* towards seeing it *as an argument about the appropriateness of the inferences drawn from the results* and the consequences of these inferences. Even a well-constructed test becomes invalid if results are misunderstood or misused.

Initially validity was treated in terms of correlating test scores with an external criterion (Thurstone, 1932), for example school grades, in order to generate a predictive validity coefficient as an empirical index of the degree to which the test measured what it claimed to measure. This ignored the content of the test and the purpose for which it was used, so the reaction to this was to place more emphasis on content validity (Cureton, 1951). The next step was to recognise that content was shaped by the construct being assessed. By the time of the first American Psychological Association guidelines in 1954 (APA, 1954) four types of validity were in play: predictive, content, construct, and concurrent (how the assessment correlated with similar tests).

This was gradually overtaken by the recognition that validity was not simply a static property of a test, which was the way it was (and often still is) presented in test manuals, but depended also on how the test was used. A well-constructed test can still be used inappropriately; we would recognise immediately that to use a maths test as the sole selection mechanism for art school would be invalid. The 1966 *APA Standards* incorporated the *use* of the test as a validity concern. This logic continued to be worked out by Cronbach (1971) who was moving towards test validation as 'a comprehensive, integrated evaluation of the test' which was centred on construct theory and the impact of the results.

These theoretical developments culminated in Samuel Messick's seminal 1989 paper. His definition of validity has become the basis for subsequent theorising:

¹ The 2008 Code of Practice, the first from Ofqual, is more likely to have validity as an organising principle. Cambridge Assessment is also currently making its approach to validity more explicit.

Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (p.13)

The 1999 *AERA Standards* distil this nicely: “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (AERA, 1999, p.9)

Bound up in Messick’s definition was the notion of social values and consequences:

For a fully unified view of validity, it must also be recognised that the appropriateness, meaningfulness and usefulness of score-based inferences depend as well on the social consequences of the testing. Therefore, social values cannot be ignored in considerations of validity. (Messick, 1989, p.19)

There has been support for the ‘social consequences’ position in both America (Shepard, 1997; Linn, 1997) and in England (William, 1993; Gipps 1994), though it has been challenged by those who regard it as ‘a step too far’. Critics such as Popham (1997) and Mehrens (1997) argue that while the impact of test usage should be evaluated, this should be kept separate from validity arguments. This debate has recently been reopened by Lissitz and Samuelson (2007) who want to simplify validity by representing it as content validity, with all concerns about the utility of test use relegated to separate evaluation. This has been responded to by such as Mislevy (2007) and Kane (2008) who see it as a regressive move.

For present purposes I take the key messages from current validity theorising to be:

1. Validity is based on the purpose(s) of a test and how effectively the interpretation and use of the results serve each purpose.
2. It is concerned with how effectively the construct/domain being assessed is being sampled.
3. Reliability and fairness are part of validity. Unreliable results undermine confidence in their interpretation, unfairness will also lead to misinterpretation of the results.
4. The impact of test results, on teaching and learning and on how learners see themselves or are seen by others, is part of any validity inquiry.

My task now is to apply these understandings to ability testing.

Ability testing

What is ability testing? The American *Standards* define it as:

The use of standardized tests to evaluate the current performance of a person in some defined domain of cognitive, psychomotor, or physical functioning (1999, p 171)

This is a careful definition which reflects the *Standards’* concern with validity, both in terms of the construct being sampled and what is inferred from the results. Ability is demonstrated through current performance, so it is about attainment in a particular domain. What distinguishes it from *achievement testing* is that the latter involves ‘a content domain in which the test taker had received instruction’ (*ibid*, p171), the assumption being that ability tests are more generalised and less content-based. The *Standards* do not distinguish between aptitude and ability. This has benefits, as can be seen from England where selection of up to ten per cent of the entry for specialist schools (which nearly all secondary schools are) can be based on aptitude testing, while ability testing is not permitted because

of its emotive IQ legacy. The chief adjudicator of selection appeals has commented “One of the difficulties is that the law uses these two words as if they were separate things and actually they are not”².

So far so good. I see this as a level 1 definition which should limit interpretation to current performance. You have ability in music because you can play or sing, you have ability in maths because you can solve mathematical problems³.

The slide into level 2 interpretation (ability as the cause, rather than the outcome, of learning) begins with the predictive claims. Not only does an ability test tell you about current functioning, it can also predict future functioning. This can either be interpreted as current functioning provides a good indicator of future performance (those doing well at 11 will do well at 16), a level 1 interpretation, or more is inferred.

It is at this point that ‘excess meaning’ begins to feature. Michael Kane in his recent authoritative chapter on validity in *Educational Measurement* (2006) argues:

Many trait labels have rich associations and are highly value laden, especially if they have been part of the language for centuries. In adopting a pre-existing trait label, a test developer is implicitly adopting this excess meaning as part of the proposed interpretation or is taking on an obligation to counteract any unwarranted inferences based on the trait label. (p.34)

What this means is that to test ability in an English-speaking context risks activating its historical significance which links it to views of intelligence and intelligence testing. David Gillborn and Deborah Udell (2001) have called ability testing *the new IQism* since ability “has come to be understood (by policymakers and practitioners alike) as a proxy for common sense notions of ‘intelligence’” (p.65). As one of their teachers put it “you can’t *give* someone ability” (p.78).

They point out the paradox that teachers who would shun the use of IQ tests appear quite happy to use ability tests which predict subsequent performance. They argue that

...it acts as an *unrecognised* version of ‘intelligence’ and ‘IQ’. If we were to substitute ‘IQ’ for ‘ability’ many alarm bells would ring that currently remain silent because ‘ability’ acts as an untainted yet powerful reconstitution of all the beliefs previously wrapped up terms such as intelligence. (p.81)

So my concern is that while ability could simply be an alternative for ‘achievement’ or ‘attainment’, the reality is that it shares the assumptions of intelligence testing: that ability is seen as *the cause of achievement, rather than a form of it*. The tendency is then to treat this as a fixed endowment with which children arrive at school. So an ability score has much the same power as an IQ score to shape learner identities (‘low ability’) and determine teacher expectations (Stobart, 2008).

This is a theme which has been picked up by Susan Hart and colleagues in their *Learning without Limits* research project. Their concern was that ability labelling “exerts an active, powerful force within school and classroom processes, helping to create the very disparities of achievement that it purports to explain” (p.21). In response they set about creating learning settings which assumed ‘transformability’ rather than fixed ability.

² *Times Educational Supplement (TES)*, 11.8.2006, p.4. As if this was not messy enough, schools also include prior achievement in these procedures, leaving him to comment that distinguishing between aptitude and attainment was “the sort of exercise lexicographers get up to when they haven’t enough to do”

³ This also seems to be the technical definition of it in Rasch modelling where one of the parameters is ‘ability’ and simply refers to an individual’s score rather than making any further assumptions.

The idea that ability measures become self-fulfilling ties in well with James Flynn's (2007) notion of 'individual multipliers'. His idea, related to IQ testing, is that small initial differences are systematically increased through interactions in the environment. For example, you and I are about the same at basketball but you are slightly taller than me so you are picked for the basketball group. You get coaching there, so get better, and are selected for the team, where you get even better. Fairly soon, as a result of more training and practice, there will be talk of 'natural ability' – a classic level 2 interpretation. Meanwhile I am seen as having only limited ability and only play socially⁴. Now transfer this logic to you passing the 11+ (the intelligence test used for secondary selection) by one mark and me failing it by one and watch the multipliers activate from that moment onwards.

So what's the problem?

This may seem like a trivial matter – an interpretation of ability which has echoes of IQ testing and British and American views of fixed and inherited intelligence. My argument is that there are serious consequences for learner identity and for teaching and learning which stem from the 'excess meaning' implicit in ability testing. Schools are full of ability talk – we have streaming by ability, mixed ability, gifted and talented classes for those at the top of the ability range. We even have children labelled as 'over-achieving' because their examination results were better than their ability test (taken five years earlier) predicted.

This level of interpretation is a serious threat to the validity of ability tests. Like IQ tests, which many strongly resemble, there is a level of inference that the tests cannot support. What is happening is that we are moving from a test of general educational achievement to make inferences about something the learner possesses before school (possibly at birth), the *cause* of learning.

How does this come about? The philosopher of science Ian Hacking has developed a broader argument about how "sometimes our sciences create kinds of people that in a sense did not exist before. This is making up people" (p.2)¹. His own examples include Multiple Personalities and obesity. The mechanisms by which these socially created classifications are brought into being are particularly relevant to my arguments about intelligence and ability testing. Hacking describes these as ten *engines of discovery* that drive this process: 1. Count, 2. Quantify, 3. Create Norms, 4. Correlate, 5. Medicalise, 6. Biologise, 7. Geneticise, 8. Normalise, 9. Bureaucratise, 10. Reclaim our identity (p.10). The development of IQ testing followed precisely this trajectory, even to the extent of the early IQ testers creating new statistical techniques (for example scaling, normal distribution and correlational techniques) to develop engines 1-4. IQ was then *biologised* and *geneticised* by giving it a physiological basis and treating it as largely inherited. This was then built into schooling provision (engines 8 & 9), for example, 11+ selection in the UK. The resistance came with the social recognition of the unfairness of this form of selection.

What the 'engines of discovery' illustrate is how we can add layers of meaning to a construct which rely on increasingly elaborate interpretations of the assessment evidence. Alfred Binet, the French founder of intelligence testing managed to stay at level 1 and castigated the 'brutal pessimism' of those who extrapolated from the results that intelligence was fixed and innate. For him the job of education was to improve intelligence - something which the leading US and British IQ testers did not think possible. Are there some contradictions in the way we use ability – we can improve a specific ability yet the underlying ability is constant?

⁴ Michael Howe takes a similar approach in his 'talent account': 'prodigies have almost always received very considerable help and encouragement prior to the time at which their ability has been seen to be remarkable' (p.132). His book *Genius Explained* (1999) develops this by looking at the biographies of recognised geniuses.

Who does ability testing anyway?

One response to all this may be to argue that it is a fuss about nothing because assessment has moved to testing curriculum-based achievement rather than ability. This would certainly be a mistake in England and America, not because there is not plenty of achievement testing, but because the Cognitive Abilities Test (CAT) plays a key role in most secondary schools in England and the SAT remains a key selection test in college admissions in the US. Add to this the reliance on ability tests in occupational selection, the use of IQ tests in selection at 11+ (in parts of England and in Northern Ireland) and SAT-like tests in other countries (eg the SweSat in Sweden) and it can be seen that this is a far from marginal issue.

Case study 1 The Cognitive Abilities Test (NFER Nelson)

The Cognitive Abilities Test is now in its third edition. There are six overlapping levels (A-F) which cover the age range 8-15 years. It 'assesses an individual's ability to reason with and manipulate different types of symbols' (p1) These types include words, quantities and spatial, geometric or figural patterns, each with its own fixed-choice, timed test.

For present purposes my interest is in the claims made for the test and how results are interpreted, rather than in the test details themselves. The third edition has tried to distance itself from the 1984 version by reducing the dependence on achievement (eg the Vocabulary test was dropped as 'it is primarily an achievement test' – interesting given some IQ tests still retain it). What is assumed is that 'All children educated in UK schools and exposed to modern cultural influences should have had an opportunity to acquire the background knowledge needed to answer the questions' (p.1).

The combination of verbal and quantitative reasoning constitute 'academic ability' (p.2) whereas the Non-Verbal Battery 'measures what has been termed 'fluid intelligence', that is, an ability to reason that is not strongly influenced by cultural and educational background' (p.2). This is somewhat undermined by the later warning that 'Caution will need to be exercised when interpreting low scores if the pupil concerned comes from a non-western cultural background as he or she may not have experienced these types of activities before' (p.49). (See Flynn (2007), Stobart (2008) for discussions of 'culture-reduced' testing).

We have seen that validity is about the inferences drawn from the results rather than simply about the quality of the test. The CAT3 is careful in its interpretation of the test norms and offers the caution:

First, it should be clearly recognised that CAT3 measures developed abilities....[These] represent the interaction of a life history of experiences impinging on a specific biological organism. But the dependence of test scores on experience does not negate the value of the test in helping to understand the individual as he or she is at the present time'. (pp.52-3)

This is good level 1 reasoning: scores telling us something about current attainment and recognising the role of experience. So do I have a problem with the CAT? Yes, because the consequence of the use of 'ability' is that the excess meaning it carries changes the interpretation of scores in many schools and policy contexts. The first step in this are the predictive claims; the CAT for 11 year olds, taken on entry into secondary schools will provide predictions of what pupils will achieve at 16 in national examinations (the GCSE). The CAT score often becomes in practice a fixed measure of ability which is used to track progress and predict outcomes (with all sorts of whizz-bang charts to help). It is then a relatively easy step to talk about under-achieving and over-achieving relative to a pupil's potential. At this point the cultural legacy of beliefs about fixed and innate intelligence, which are never far below the surface in Britain, kick in.

Am I overstating the case? Probably, but take this example from government- funded Teacher Training Resource Bank report of research on *Raising students' performance in relation to NFER CAT⁵ scores*. This study compares 'the responses of under-achievers (those who do not do well in relation to their measured ability) and over-achievers' (who do well...)(p.1) in order to reveal the motivational factors that may account for this. For the record, over-achievers, in contrast to underachievers, 'saw learning was for their benefit and satisfaction; treated other students with sensitivity and respect; listened to the ideas of other students...worked with teachers in a polite way and attended all their lessons regularly' (summary report p.2).

What would I do about this? Well, a name change might help reduce excess meaning (the Cognitive Reasoning Test? The Generalised Achievement Test?). Before this is declared impossible let us consider the SAT.

Case study 2. The redefining of the SAT

The SAT is the American the college entrance test taken by millions. For most of its eighty year history, it was called the *Scholastic Aptitude Test*. It has recently dropped this full title because 'aptitude' was proving problematic. This was because it is clearly not a 'school-free' test of potential, but rather a measure of general educability based on language and mathematical skills, themselves dependent on opportunity and schooling. It also helped resolve the paradox of courses being offered to improve performance (thus increasing one's aptitude). This is now explicitly acknowledged:

It tests students' knowledge of subjects that are necessary for college success: reading, writing, and mathematics. The SAT assesses the critical thinking skills students need for academic success in college—skills that students learned in high school. (CollegeBoard p.1)

The Educational Testing Service (ETS) now treats SAT as an empty acronym – SAT does not stand for anything.

So ETS has explicitly moved back to a level 1 interpretation. This resonates with the words of the remarkable Carl Brigham who first introduced the test (an adaption of the Army intelligence test) in 1926, but soon after recanted on the claims that were being for it:

The more I work in this field, the more I am convinced that psychologists have sinned greatly in sliding easily from the name of the test to the function or trait measured...I feel we should all stop naming tests and saying what they measure...if we are to proceed beyond the stage of a psycho-phrenology (1929 – in Lemann, 2000, p.33)

Where does this leave us?

Current formulations of validity see it as an evaluation of how well the interpretation of test results serves the intended purpose. The bigger the claim for the purpose, the more complex the validity argument will be. This interpretation of results involves those who use them and their impact.

My concern with ability tests is that, while the test developers may make relatively modest (level 1) claims, the use of 'ability', with all its excess meaning in much of the English-speaking world, leaves those who use them in schools reading far too much into them. My ability level is fixed, and my subsequent achievements will be judged in relation to this. If I work hard and am motivated I may even 'over-achieve'. Ability interpretations have multiplier effects on learners and their sense of identity – a good score may put you in the 'gifted and talented' stream with extra resources and teaching. A low score means you have 'low ability'

⁵ <http://www.ttrb.ac.uk/aboutUs.aspx?menuId=1173>

- not a term used by the test developers, but widely used in schools. And, despite test developer warnings that 'all measurement contains error' and 'a difference between two test scores [for the same person] is fairly common' (CATS3 p54), your score will be treated as fact.

What are the responsibilities of test developers in such over-interpretations? The idea that validity incorporates consequences means that test developers cannot wash their hands of misinterpretation, particularly widespread ones. This is true of any test, not just those of ability. ETS took the step of de-naming the SAT because of such problems, though these do not automatically go away if the historical legacy remains.

My fundamental concern is that the interpretation of ability scores has become counter-productive in terms of learning and learner identity. David Olson's critique of such assessments is that they take away personal responsibility for our learning, since we are not responsible for our dispositions or abilities:

One acquires competencies as one takes on responsibility for certain standards...One is not responsible for mere dispositions, dispositions are causes of actions not reasons for acting. A theory of education has to spell out how children take on responsibilities for learning and how one, whether teacher or learner, goes about judging whether those responsibilities have been met....Teachers are notoriously disposed to explain children's success in terms of putative abilities and learning styles rather than on the conditions that make learning easy or difficult. (2006, p.42)

This means accepting more responsibility for our learning: '...agency, intentionality and responsibility could become the central features of a psychology that has special relevance for education. Abilities, traits and dispositions can be left to find a new place in the natural sciences or else relegated to the dust-bin of history' (Olson, 2006, p.43).

To do this we may to address the current discourse of ability, and as test developers we have a responsibility in this, Kane's 'taking on an obligation to counteract any unwarranted inferences based on the trait label'. A first step may be to look for alternative terms:

Sometimes an expression has to be withdrawn from language and sent for cleaning – then it can be put back into circulation. (Wittgenstein)

What shall we wear in the meantime?

References

- AERA, APA & NCME (1999) *Standards for Educational and Psychological Testing* Washington, DC, American Educational Research Association.
- APA (1954) *Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal*, Washington, DC, American Psychological Association.
- APA (1966) *Standards for Educational and Psychological Tests and Manuals*, Washington, DC, American Psychological Association.
- APA (1974) *Standards for Educational and Psychological Tests and Manuals*, (Washington, DC, American Psychological Association).
- CollegeBoard (2006) *SAT Reasoning Test*
<http://www.collegeboard.com/student/testing/sat/about/SATI.html> accessed July 2008.
- Cognitive Abilities Test 3 (undated) *Administration Manual*, London, NferNelson.
- Cronbach, L. J. (1971) Test Validation, in: R.L. Thorndike (Ed.) *Educational Measurement, 2nd ed.* (Washington DC, American Council on Education), 443-507.
- Cureton, E. E. (1951) Validity, in: E.F. Lindquist (Ed.) *Educational Measurement, 1st ed.* (Washington DC, American Council on Education), 621-694.
- Flynn, J. R. (2007) *What is Intelligence*, Cambridge, Cambridge University Press.
- Gillborn, D. & Youdell, D. (2001) 'The New IQism: Intelligence, "Ability" and the Rationing of Education', in J. Demaine (Ed.) *Sociology of Education Today*. Basingstoke: Palgrave.
- Gipps, C.V. (1994) *Beyond Testing: toward a theory of educational assessment* (London, Falmer Press).
- Hacking, I. (2006) *Kinds of People: Moving Targets*, The Tenth British Academy Lecture. Online. Available <http://www.britac.ac.uk> (accessed 1 July 2006).
- Hart, S., Dixon, A., Drummond, M.J. and McIntyre, D. (2004) *Learning without Limits*, Maidenhead, Open University Press.
- Howe, M. J. A (1997) *IQ in Question*, London, Sage.
- Kane, M.T.(2008) Terminology, Emphasis and Utility in Validation, *Educational Researcher*, 37, 2, 76-82.
- Kane, M. T. (2006) Validation, in R. L. Brennan (Ed.) *Educational Measurement* 4th edition, American Council on Education, Praegar.
- Kane, M. T., Crooks, T. J., & Cohen, A. S.(1999) Validating measures of performance. *Educational Measurement: Issues and Practice*, 18, 2, 5-17.
- Lemann, N. (2000) *The Big Test*, New York, Farrar, Strauss & Giroux.
- Linn, R. L. (1997) Evaluating the Validity of Assessments: The Consequences of Use, *Educational Measurement: Issues and Practice*, 16, 2, 14-16.
- Lissitz, R. W. & Samuelson, K. (2007) A Suggested Change in Terminology and Emphasis Regarding Validity in Education, *Educational Researcher*, 36, 8, 437-448

- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16, 2, 16-18.
- Messick, S. (1989). Validity, in: R. L. Linn (Ed.), *Educational measurement 3rd ed.* (New York, American Council on Education), 13-103.
- Mislevy, R. J. (2007) Validity by Design, *Educational Researcher*, 36, 8, 463-469.
- Popham, W. J. (1997). Consequential validity: Right concern – wrong concept, *Educational Measurement: Issues and Practice*, 16, 2, 9-13.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16, 2, 5-8.
- Stobart, G. (2008) *Testing Times*, Abingdon, Routledge.
- Thurstone, L. L. (1932) *The Reliability and Validity of Tests* (Ann Arbor, Michigan, Edwards Brothers).
- William, D. (1993) Validity, dependability and reliability in national curriculum assessment, *The Curriculum Journal*, 4, 335-350.
- Wood, R. (1991) *Assessment and Testing*, Cambridge, Cambridge University Press.
-