

The importance of being valid

Alastair Pollitt & Ayesha Ahmed

Cambridge Exam Research

www.camexam.co.uk

**A paper presented at the 10th Annual Conference of the
Association for Educational Assessment - Europe**

Malta, November 2009

Introduction

For many years we have worked with professionals involved in various ways with the activity of educational assessment. Our role has included helping with the clarification of the purposes of assessments, the writing and use of test specifications and, in particular, with the creation of test items. We have focused on developing models to help question writers produce tasks that will measure the knowledge and skills that they want their examination to assess, and to develop marking schemes that will reward students for high levels of achievement in them. We have studied the mental processes of the students who attempt these tasks, and of the assessors who create and evaluate them.

As this programme has progressed, we have become increasingly dissatisfied with the generally accepted doctrine of validity. We believe that the validity of any assessment enterprise depends crucially on what happens when a student's mind meets an assessment task, and when the outcome of that meeting is evaluated. Other aspects of the procedure are important, of course, but it is in these mental activities that validity is primarily achieved or lost.

The broadening of the concept of validity in recent decades has brought many benefits to educational assessment, but we believe that it has also confused many of the professionals we have worked with, and distracted them from their proper concerns as participants in a collaborative activity. To remedy this, we propose in this paper a simple dynamic model of validity as a quantitative property of the procedure: a model which shows that every participant is responsible – in part – for validity, and suggests how each of them can optimise their own particular contribution to the common objective of maximising validity.

The current doctrine

The concept of validity has been redefined many times (Wolming & Wikström, 2006). Early definitions were simple, or empirical, and an important step forward was the formal definition of *construct validity* by Cronbach & Meehl (1955) to capture the idea that a test was valid to the extent that it measured a hypothesised psychological trait – evidence for this would come from many sources, each giving some support to the inference that the test measured what it was intended to measure. But other notions, such as *content*, *criterion*, or *face* validity were prominent, and each reflected different desirable features of a test or examination.

Messick (1989) laid out a framework for considering all of these concepts in a *unitary validity framework*, in which construct validity played a central role, but where the other aspects of quality contribute: he defines validity as “an evaluative summary of both the evidence for and the actual—as well as potential—consequences of score interpretation and use”. A key feature of the current doctrine is a shift from the concept of ‘validity’ to that of ‘validation’, the process of gathering evidence to support the inferences that a user of test results wants to make.

This expansion of the concept has extended the responsibility for validity to those who use the results of tests. Every decision made as a result of an assessment can be challenged in a quasi-legal way: can you justify the decision you took? The change has brought benefits, notably in emphasising the importance of purpose, that a test must be ‘fit for purpose’, and in reminding everyone involved that any assessment results can be misused if the original purpose is ignored.

But it has brought problems too. Our work with assessors has led us to believe that the central concept of validity – measuring what you want to measure – is in danger of being lost by this emphasis on interpretation of results instead of on the test itself, and that the central procedure of test construction is becoming devalued.

Standards for Educational and Psychological Testing

Consider the authoritative ‘Standards’ document published by APA, AERA and NCME (Standards, 1999). Its first chapter, *Validity*, begins with this sentence:

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.

It continues:

It is the interpretations of test scores that are evaluated, not the test itself.
(‘Standards’, p9)

It is perhaps no surprise, then, that in this chapter not one of the ‘standards’ refers to the theoretical basis of test construction or item writing; when they address test constructors at all it is only in terms of the empirical evidence they should provide of relationships between scores and other external measures, or amongst the scores themselves – after the test has been used. The process of designing the test items to ensure that they actually require the kinds of thinking that the test is supposed to measure is addressed in just one out of the twenty-four ‘standards’:

Standard 1.8

If the rationale for a test use or score interpretation depends on premises about the psychological or cognitive operations used by examinees, then theoretical or empirical evidence in support of these premises should be provided.

We find it difficult to accept that *any* test, or the interpretation of *any* test scores, can be valid if the students’ minds were *not* doing the things the item writers intended them to do (Pollitt & Ahmed, 2008).

In contrast, the second chapter – *Reliability and errors of measurement* – contains twenty ‘standards’, of which at least eighteen are directed to test constructors. The implication, whether intended or not, is that test constructors – the people who design the assessment and write the questions and mark schemes – should concern themselves mainly with reliability, and leave validity to those who come after them. There is little encouragement for test developers who are struggling to design a better way of measuring something, when they are reminded that the test users are liable to misuse their test however brilliantly it is created.

As a consequence of the perspective presented by the *Standards*, it is now commonly stated that it is wrong even to talk of the validity of a test. We believe this to be profoundly misguided, and harmful. Not only does it make sense to talk of a valid test, but it makes even more sense to talk of the validity of a test item[§].

A valid test is not a “category error”

Several authors have claimed that it is a “category error” to talk of a test as valid (eg Wiliam, 2008), meaning to imply that ‘validity’ is a characteristic that cannot logically be applied to a test. The argument derives from Gilbert Ryle’s introduction of the term

[§] We use the term ‘item’ here to refer to the combination of a test question and its mark scheme: together they constitute the fundamental unit of assessment.

‘category mistake’ to describe the Cartesian doctrine that ‘mind’ and ‘body’ can be talked of in similar ways, that both are locatable ‘objects’ in the world; by analogy it is claimed that ‘tests’ and ‘the interpretations of test results’ do not belong to the same logical category and to apply ‘valid’ to the former is a category error. This, however, is not a valid argument.

Ryle defined a ‘category mistake’ in *The concept of mind* (Ryle, 1949) with a series of examples:

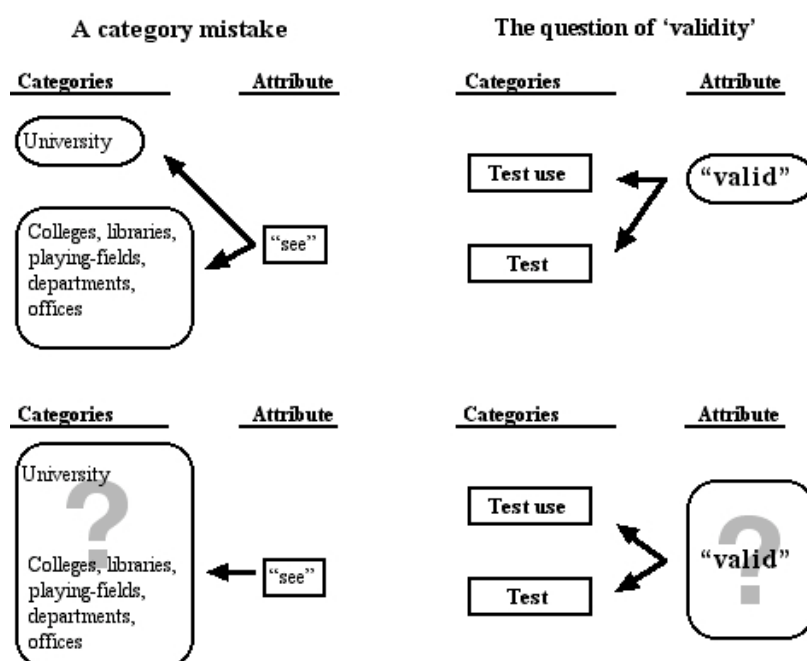
the foreigner who, having been shown the colleges, libraries, playing fields, museums, scientific departments and administrative offices of Oxford or Cambridge, then asks ‘But where is the **university**? I ... have not yet seen the **university**...’;

the child who saw a parade of battalions, batteries and squadrons and asked when the **division** was going to appear;

.....

asking **where** a person’s mind is.

In all of these the mistake is a real confusion of categories – the University, the division, or the mind are wrongly seen as belonging to a category for which the simple attributes ‘see’, ‘appear’ or ‘where’ are appropriate. But in our field there is no confusion of categories. Everyone can distinguish a ‘test’ from ‘the interpretation of test results’.



The point at issue concerns the word *valid*, not the categories of *test* and *test use*. It is whether or not the word *valid* can properly be used to describe an object as well as to describe the use of that object. There is no category mistake here, nor any other profound logical principle, but merely an argument about the meaning of a word.

Real world meaning of valid

What then does valid ‘really’ mean? Concordance dictionaries, which indicate the relative frequencies of different uses of a word, reveal a surprising pattern for ‘valid’. The Collins Cobuild corpora of written English (British and American) give the following relative frequencies:

Rank	Concordance	Relative frequency	Rank	Concordance	Relative frequency
1	offer	102	8	claim	14
2	ticket	32	9	entry	12
3	licence	25	10	card	12
4	certificate	25	11	conditions	10
5	passport	19	12	coupon	9
6	law	15	13	sales	8
7	voucher	14	14	documents	8
			15	discount	8

By far the commonest use of the word is in a phrase like “Offer valid until 31st...”. Then come a series of phrases concerning valid documents of one sort or another. The only example of the use of valid as a logical term in the top fifteen is ‘valid claim’, though ‘reason’ and ‘principle’ are just a little further down the list.

The number of hits in a Google search provides a rough, but up-to-date, indication of the same pattern. Here are the numbers of hits for selected collocates:

"valid offer"	2,375,100	(inc. “offer valid”)
“valid claim”	943,000	
"valid passport"	724,000	
"valid question"	358,000	
"valid test"	267,800	
“valid licence”	181,300	
“valid ticket”	171,800	
“valid interpretation”	56,960	
“valid inference”	49,320	

Note how common it is to see both ‘valid question’ and ‘valid test’.

Dictionaries usually distinguish about three senses for ‘validate’. For example, Chambers 20th Century Dictionary offers:

1. to make valid;
2. to ratify;
3. to confirm, substantiate

and The Free Dictionary by Farlex gives:

1. To declare or make legally valid.
2. To mark with an indication of official sanction.
3. To establish the soundness of; corroborate.

It is clear that, to most people, ‘valid’ usually concerns the first two of these, concerning the legal or official status of documents. In particular, there is an important difference between meanings 1 and 3: the first refers to *making* something valid while the third refers to *checking* its validity after the event. This is the distinction that we want to hold to in this paper. Our concern in the field of assessment is that the current doctrine of validity and validation is mainly about Meaning 3 – checking validity – and not enough about Meaning 1 – creating validity in the first place.

The student/item interaction

Modern test theories, especially latent trait ones, make explicit that the process of measurement takes place when a student's mind meets an item. The simplest Rasch model equation shows this clearly:

$$\log \text{ odds} = \beta_n - \delta_i$$

where β_n relates to the ability of the student and δ_i relates to the difficulty of the question. Here is the essential interaction that determines an assessment: a student is measured by a question, and this event is repeated for every question that the student faces. It follows that this is the most important step of all, for if this interaction goes wrong the essence of the assessment is faulty; how then can we claim validity?

This interaction is what test constructors are concerned with when they conceive of and write the questions and mark schemes that they hope will produce valid assessment. Our work with question writers has led us to express it in the requirement that an item will only be valid if the mental processes that we want to measure are indeed what decide how many marks a student gets in answering each and every question. Borsboom (2005: 160) shares this view, writing in his development of the concept of validity: "What is meant is that traits cause observed scores".

To focus on this causal link between the mental processes of the students and the scores they get, we are in effect led back to a theoretical consideration of the trait, to what we understand to be the central question for every assessment: what is it that we *want* to measure?

Importance

Let's begin with a "widely used definition of assessment":

The process of seeking and interpreting evidence for use by ...
(ARG 2002, pp. 2–3, quoted by Stobart, 2009, p5)

The full quotation refers to formative assessment, but we can take this first part of the sentence as applying to any form of assessment. The idea is always that assessment is a process, that *evidence* is to be sought and interpreted for a purpose. So: what evidence do we want to seek and interpret? Evidence of what?

Our answer is that we seek evidence of what is *Important*. This may seem a trivial reply, but the point is that it is possible to use a statement of what is important as a focus for ensuring validity in the interactions between student and questions. Indeed, such a statement can act as a focus for more than just assessment: a statement declaring what is important can be used by those who build a curriculum, by the writers of questions and mark schemes, by the markers and graders, and even by the parents, employers, selectors and politicians who observe the educational system from the outside. For all these participants and observers, constant reference to the statement of what is important will help to maximise the validity of the whole system.

Our interest is with the assessors. Questions should be written to test what is said to be important in this statement; and the marking or judging schemes should evaluate students' responses in terms of the evidence they show of what is important.

The importance statement

In England, the Qualification and Curriculum Authority published support materials on its web site for each subject in the country's National Curriculum. Every subject's section begins with an Importance Statement, and the Geography one is quoted here:

The importance of geography

The study of geography stimulates an interest in and a sense of wonder about places. It helps young people make sense of a complex and dynamically changing world. It explains where places are, how places and landscapes are formed, how people and their environment interact, and how a diverse range of economies, societies and environments are interconnected. It builds on pupils' own experiences to investigate places at all scales, from the personal to the global.

Geographical enquiry encourages questioning, investigation and critical thinking about issues affecting the world and people's lives, now and in the future. Fieldwork is an essential element of this. Pupils learn to think spatially and use maps, visual images and new technologies, including geographical information systems (GIS), to obtain, present and analyse information.

Geography inspires pupils to become global citizens by exploring their own place in the world, their values and their responsibilities to other people, to the environment and to the sustainability of the planet.

In just 160 words it aims to capture the essence of why students should study geography, and it gives anyone who reads it a good idea of what would distinguish a 'good' geography student from a 'poor' one.

The Importance Statement has many uses. Once it has been accepted by all the relevant stakeholders, it provides a warrant for designing the syllabus, for choosing what must be included and as a basis for making the difficult decisions when there is too little time to include everything that everyone would like to see in it. Parents and prospective pupils may find it more useful than the detailed syllabus when deciding which subjects a child should study, and professionals outside the school system can see what knowledge and skills the course is meant to develop in the pupils.

For teachers it provides a general criterion that they can use to judge the value of any activities they might consider using with a class, a constant reminder of what they are meant to achieve with their pupils, and a simple statement of what the examinations or tests will deem important.

Assessors, it follows, must also use the Importance Statement in their work. They must ensure that their tests do indeed seek evidence of what is ***agreed to be important***, and that they do indeed give credit for evidence that students' have learned what the Importance Statement defines as important. For this assessment function, the brevity of the statement is important; there are other documents that assessors might use to guide them in creating tasks and marking responses – aims, objectives, learning outcomes, grade descriptors – but these are often too detailed, and too list-like, to keep the assessor's mind focused on what the overall purpose of the test is – measuring what is important.

The Importance Statement expresses the general principles for judging importance, and is brief enough for every participant to keep it clearly in mind, but also has enough detail to be useful in practice. Assessors will use it constantly when writing questions, when constructing mark schemes, when scoring responses, and when setting grade boundaries and standards. The essence of validity is that all assessors pay proper attention to the

agreed statement of what really matters at each and every step: a good Importance Statement will ensure that validity is maximised at the start of the assessment procedure.

Validity

The preceding discussion requires us to propose a model of validity that is useful to everyone involved in the educational process, from the conception of the curriculum to the use of certificate examination results by selectors and employers. As a metaphor for such a model, consider a *Bucket Brigade*, the human chain that sometimes transports water from where it is plentiful to where it is needed.

The Bucket Brigade model for Validity

Imagine a village house is on fire. A hundred metres away is a well, but there is no pump to bring the water to the house. The traditional solution to this is the *bucket brigade*: every household contributes a bucket and a person to help fetch water. Rather than everyone running with their own bucket, from well to house, they carry it only a short distance then hand it on to the next person. The aim is to get as much water to the house as possible, as quickly as possible, and it is important that each person does what they can to keep the water in the bucket without wasting too much time to do it.

This seems appropriate as a model for validity. When an assessment is first imagined, the assessors involved will have a notion of the ideal test they would like to create, reflecting very closely the statement of what is considered important in studying that subject. They will have considered with care what would constitute a sufficient sample of evidence of the achievement they hope to see; the Importance Statement is the well at which they fill their bucket of potential validity as full as they can. This is the crucial first step in developing a valid assessment: the job of the assessment is to elicit and evaluate evidence of what is important, and that evidence is derived from the Importance Statement.

From here on, validity can only be lost. As they proceed to specify the test in detail compromises must be made, and the test will not be as good – that is, valid – as they had hoped. When they pass the bucket on to the item writers it will not be as full as when they first imagined it. The question writers in turn will create mark schemes and tasks to elicit and evaluate evidence of how well students have achieved what's considered important, but these tasks will not do the job quite as well or as completely as the writers hoped – and some more validity will splash away.

Then some students will mis-interpret some of the questions, some markers will mis-interpret their responses or even the mark schemes, and so on. At every step in the process, a little more, or a lot more, validity will be lost. At every step it is the responsibility of every participant to try to maintain as much validity as they can.

The Entropy law of validity

One important difference between this model of validity and the current doctrine can be stated in what we call the Entropy Law of Validity:

During the assessment process, validity always decreases.

The Bucket Brigade model makes it clear that those who plan the assessment must design as much validity into it as possible, as the carrying forward of that plan cannot add new validity, but can at best preserve most of what's there as compromises are inevitably made. Another formulation of the law is that:

Validity can only be lost, it can never be increased.

Everyone, from the start of the process right through to the final interpretation of the results, must see their duty as to maintain as much validity as they can. We must not argue that validity is only about the interpretation of the results of assessment, since the price of not paying enough attention to validity before and during the test itself may well be that there is just not enough validity left in the bucket for the results to be interpreted with any confidence at all. Validity is a precious commodity that is invested in the test when it is first imagined, and that must be conserved as well as possible throughout the process. In the next section we will outline some of the requirements for maintaining validity.

How to maintain validity

1 Writing the Importance Statement

It is essential that we fill the bucket as full as possible at the beginning. The Importance Statement must be accepted by everyone involved as a reasonable description of what matters when students study a subject. For summative assessment involving certification this means that a broad representation of those with an interest in the process should be included in the drafting of the Statement.

It is evident that there are differences in different countries as to ‘Who owns the standard’: the extent to which the meaning of ‘being good at X’ is decided by the teachers who deliver the education or by external authorities. To ensure that validity can be maintained by every participant through the process, however, it is essential that teachers, assessors and the users of the results understand and endorse this statement in the same way.

2 Specifying the test

Designing an assessment is always a matter of making compromises. The resources available for the test are never enough to support the best procedures that the assessors would like to use. Demands for security, for reliability, or the limits of time, money or technology, or the need to minimise potential biases will all reduce the amount of validity that can eventually be built into the process.

Validity in this sense is difficult to measure, since there are so many factors to be considered and balanced in the eventual compromise. Assessors can, however, always judge whether a given change to the test will remove more or less validity from the bucket; it is the duty of a professional assessor to judge what compromise will maintain as much as possible for the next stage.

The job of the exam is:

to award results (grades, scores) on the basis of the best possible evidence of what is important.

3 Writing the questions and mark schemes

We believe that the writing of good assessment items is a creative process that should be guided by proper attention to the Importance Statement, and that *item* means the combination of question and mark scheme, since these together constitute the basic unit of assessment. Whenever a writer forms an idea of a possible item for a test, the first step should be to decide what evidence the item might elicit of the kinds of achievement deemed important in the Statement.

We have found that the next step should be to decide what approach will be adopted for evaluating that evidence (Pollitt et al, 2008). Usually, this means choosing the best kind of mark scheme, with a focus on correctness or quality or some mixture of the two, and

choosing which demands should be designed into the task. Once it is clear what evidence is wanted and an approach to evaluation has been chosen, the question can be designed to elicit the evidence in a way that the mark scheme can handle. This apparent reversal of order – mark scheme first and then question – fits the logical sequence outlined earlier: the Importance Statement is used to determine the evidence we want to see, and we need to understand how we will evaluate that evidence before we can decide how to elicit it:

The job of the mark scheme is:

to evaluate the evidence in terms of how well it shows the students have achieved what is important.

The job of the question is:

to elicit that evidence of what is important.

To achieve this, it is important that the assessors understand how students think as they read and consider test questions. The requisite skills are an understanding of both the cognitive psychology and the psycholinguistics of students' minds, operating under stressful conditions, and a familiarity with the particular kind of students and the content material of the subject.

4 The students

This concern for understanding can be extended to include the students. An examination can be considered as a three-part communication activity: the students must understand the task they have been set by the examiners, but they must also understand the criteria that markers will use to assess their responses (Pollitt & Ahmed, 1999). This may be quite a simple issue for a multiple choice test (though even then they must know any rules about omissions and 'corrections' for guessing), but whenever a more extended response is required they must also understand what would distinguish a good answer from a poor one; they should not be in doubt – in general terms – as to what it means to show you are a 'good' student. The ability to self-assess may be considered an essential feature of education, as Sadler (2009) does: "developing evaluative expertise is a necessary but not sufficient condition for being able to produce quality works consistently". A student who has not developed this ability is not likely to perform well in an examination, but nor will a student who has developed it but does not understand the criteria for a particular task.

Low marks are fair if the students know what it means to be good but are unable to do it, or if they have not developed this ability to evaluate their own work, but they are *not* fair if they occur because the assessment criteria are kept secret from the students.

5 Marking

There are two main threats to validity during marking. The most obvious is the risk that different markers may interpret the marking instructions differently – inter-marker unreliability. Three sorts of error are identified, involving differences in severity or leniency, in range of marks awarded, or in interpretation of the trait leading to differences in rank order. The traditional approach to minimising these is to make the mark scheme ever more precise, and to train the markers to use it more consistently. But this approach may increase the second threat as it reduces the first; the urge to make the marking more objective may distort the evidence to such a degree that the resulting scores are less valid indicators of achievement.

These threats must be balanced: assessors must judge, for example, if the change in the mark scheme, or in the question, that is needed to maintain reliability risks losing too much validity from the bucket. They might consider a different approach to evaluating the

evidence if the cost seems too great. At all times, decisions should be made in terms of the constant need to assess what is Important as well and as fully as possible.

6 Grading etc

The aim of the assessors who designed the syllabus and constructed the mark scheme and the test was to generate results that validly reflect students' real achievement on what is considered Important. Any decisions about aggregating outcomes and setting grade boundaries, and about the setting and maintaining of standards, should be informed by the same concern to maintain the validity in terms of what is important.

7 Interpreting

Finally, the Importance Statement should always be available to anyone who might need to interpret the results of assessment. The current doctrine that validity resides in the interpretation of the results is not wrong: all the good work the teachers and assessors have put into maintaining as much validity as possible through to the final certification or reporting can be destroyed by a careless or ill-informed understanding of what the results really mean.

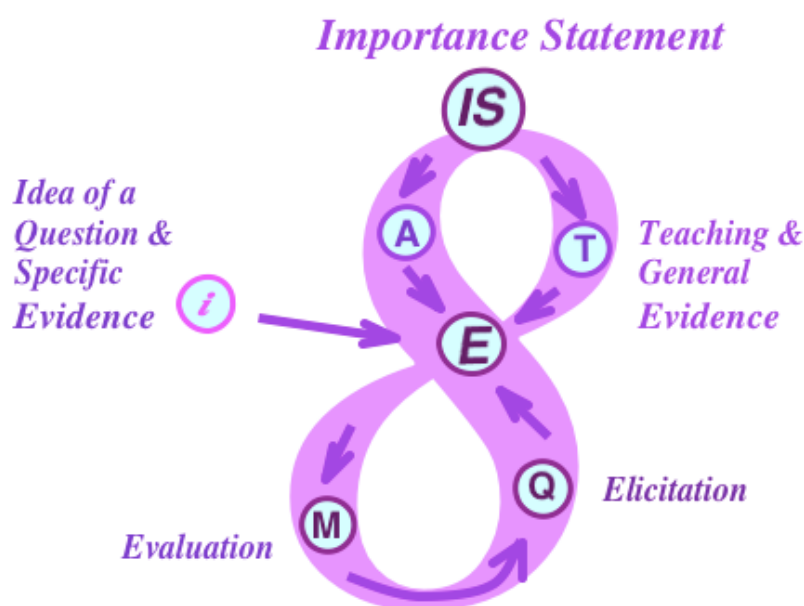
If the process so far has been done well, the results should be capable of being interpreted validly – as giving a useful measure of what the Importance Statement declared was important. Research may be used to show that the results are consistent with what would be expected of a valid test, but in education it will rarely be possible to carry out experimental studies to confirm the causal link between the trait and the scores. There will always be a temptation to interpret correlational evidence as if it were causal.

In these circumstances evidence that everyone in the process properly followed the demands of the Importance Statement, and that when every decision was made it was maintaining validity that was the principal concern, may be the best guarantee that the assessment was valid.

Summary

We summarise this argument – how the Importance Statement is the ultimate source of validity and the central role of evidence that students show they have mastered what's important – in this diagram.

The importance of evidence



The Importance Statement is the source of validity, not only for Assessment but also for Teaching. Both are focused directly on what is important in the subject being considered: teachers aim to teach what is considered important, and assessors aim to measure students' success in achieving it. For an examination, item writers think up Ideas which will allow them to elicit specific Evidence of achievement, Mark schemes which will allow them to evaluate the evidence and Questions that will elicit it.

Evidence of what is Important is the essence of Validity. The evidence must be elicited and evaluated properly, and the results of this assessment must be used with due care, but the Importance Statement is the source of the evidence and its use – the well from which validity springs.

References

- Assessment Reform Group (2002) *Assessment for Learning: 10 Principles*, University of Cambridge, UK: Assessment Reform Group.
- Borsboom, D (2005) *Measuring the Mind*. Cambridge: Cambridge University Press
- Cronbach, LJ & Meehl, PE (1955) Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Messick, S (1989). Validity. In R.L. Linn (Ed.), in *Educational Measurement* (3rd edition, page 13-103). New York: American Council on Education and Macmillan.
- Messick, S (1995) Validity of Psychological Assessment. *American Psychologist*, 50, 741-9.
- Pollitt, A & Ahmed, A (1999) *A new model of the question answering process*. IAEA Annual Conference, Bled, May.
- Pollitt, A & Ahmed, A (2008) *Outcome Space Control and Assessment*. AEA-Europe Annual Conference, Hissar, November.

- Pollitt, A, Ahmed, A, Baird, J-A, Tognolini, J & Davidson, M (2008) *Improving the quality of GCSE assessment*. Final report to QCA. London: Qualifications and Curriculum Authority.
- Ryle, G (1949) *The concept of mind*. London: Hutchinson's University Library.
- Sadler, DR (2009) *Thinking differently about assessment: Why feedback is not enough*. IAEA Annual Conference, Brisbane, September.
- Standards (1999) *Standards for Educational and Psychological Testing*. APA, AERA and NCME.
- Stobart, G (2009) *Keeping Formative Assessment Creative*. IAEA Annual Conference, Brisbane, September.
- William, D (2008) *What do you know when you know the test results?* IAEA Annual Conference, Cambridge, September.
- Wolming, S & Wikström, C (2006) *Evidence or consequence? Reflections on validity and validation*. AEA-Europe Annual Conference, Naples, November.