What can we measure?

And what should we be measuring?

Alison Wolf

King's College London

Let me start with a couple of cautionary tales and a paradox.

The state of Texas, some years ago, was horrified by stories of illiterate and innumerate teachers. The legislators decided that no-one should teach in Texas unless they met stringent standards and an eminent, external testing agency was commissioned to develop the Texas Teacher Test. Those who failed would lose their jobs. The test was developed to the highest technical standards. At the same time the Teachers' Union organised intensive study sessions for all teaching staff. In the event over 95% passed. Failures were, almost to a man or woman, either experienced vocational ('shop') teachers or teachers working with the severely disabled.

The current Labour government, some years ago, introduced performance related pay for teachers, on the assumption that this would provide an incentive for changed and improved teaching practices. Detailed criteria were drawn up: to be successful teachers had to meet these. Head teachers, who were able to judge their teachers at first hand, had responsibility for judging whether the criteria had been met. Success among those applying was virtually 100% (and almost all those eligible applied.) There was also a large increase in the testing of pupils.

The paradox involves Maths A level. In the modern UK, there are increasing rewards for those with Maths A level and with quantitative degrees, This contrasts with, for example, the 1950s and 1960s when science and engineering graduates were relatively poorly rewarded. Yet this increased demand has not found any response in the schools where maths enrolments continue to languish or decline.

Why do I start with these stories? Because, very often, we discuss assessment as an essentially technical affair – a slightly more demanding version of the measurement that goes on in a lab or on a building site. We are aware – or should be - that measurement is an imperfect business in all these environments: the signal to noise ration may be low, the confidence intervals large.

But these are seen as essentially technical concerns. They may be difficult to cope with, but difficult in the way that engineering solutions are often difficult to find. There may be measurement breakthroughs – the move from manual to electronic micrometers, for example, greatly increased the accuracy of the average reading of minute lengths. It may be harder to envisage such breakthroughs in the measurement of human traits, or mastery of knowledge, but the process is conceived as analogous.

Somewhere out there is the thing we want to measure. The better our measuring instrument, the more accurate the 'reading'. In psychometric theory, the central organising concept is that of the 'true score'. If someone could take a test an infinite number of times, without learning anything new in the process, the average score would represent their 'real' facility level – exactly as medical technicians, for example, use repeated measures to decide the 'real' dimensions of something internal.

The aim is to develop tests that are so reliable that users are confident that the 'true' and the measured score are very close – so, for example, someone taking it one day, with one marker, will gain pretty much exactly the same score as if they took it on another day with a different marker. In other words, this is a technical issue of measurement – and it is this issue of reliability that preoccupies examination and assessment authorities and their regulators in much of their daily work. Except perhaps for direct fraud in the system – which is not a common British experience - nothing threatens the credibility of an assessment system more quickly than public perception that it is deeply unreliable in the sense I have just described.

Of course, there is also the prior question of deciding what is being measured and whether this can be done directly or nor. Length – if you are not a theoretically inclined scientist or philosopher of science – is a concept that we treat as pretty robust and easy to understand, and measurement of lengths is something we do, directly,. with varying levels of accuracy or reliability. Psychological constructs such as creativity or leadership potential are, by contrast, difficult to define, let alone to measure with any degree of confidence or consensus on how to do so.

In educational assessment, we are actually in relatively easy waters. The curriculum is seen as providing a clear definition of what is to be measured; assessment questions sample the domain. Subject experts decide what the curriculum is to be, and are in fact the only real source of curriculum validity – of assurance that the domain being assessed is the 'right' one, the one we want to measure. If they also write the questions, the circle is closed. There may again be technical issues to do with clarity, question 'demand' , sampling across the domain and the like, but these are more to do with reliability than with whether the questions are 'valid' and measure what is meant to be measured.

In practice, the curriculum, as taught, is determined as much by exam questions as the questions are by the curriculum. To the extent that considerations of reliability come to dominate question-setting, this can have important repercussions for learning – a point I return to later. However, the critical point I wish to make here is rather different. It is that conceptualising assessment in this largely self-contained way, as a measurement activity, leads us to ignore an absolutely central difference between measuring human behaviour and measuring the length of a beam or the dimensions of a galaxy. Because we are measuring human behaviour the act of measurement can have profound, visible and consequential effects on that behaviour. Assessment is part of a dynamic system, it affects as well as being affected – and it affects its environment in ways which depend on how people understand and respond, consciously and purposefully, to assessment.

This is why I started my talk with two cautionary tales. They both illustrate, of course, the general law of unintended consequences, but they also illustrate two more specific points. People, unlike beams, do not simply lie there when they are being measured.

And they are particularly unlikely to do so when the measurement has specific consequences for their futures.

One of the most perceptive observations of recent years, made by the economist Charles Goodhart, should be engraved above every senior policy-maker's and every minister's desk. Goodhart's Law states that 'Any measure used for control becomes unreliable.' We can all think of recent examples in the UK public services, whether in education, where the pressure is to get as many pupils as possible to obtain'5 GCSE Grade A to C or equivalent', or to deliver a set number of 'full level 2 awards' to adults, or to cut waiting list times, or time between requesting and getting a GP's appointment. These examples, however, mostly involve the people administering a system. We need to remember that students, and candidates, and indeed patients, are active participants in the system too.

So the answer to what we can measure is '*Rather less than we like to believe.*' If we take any one component of a curriculum, whether it is Key Stage 2 English, or Part II Natural Sciences, and consider it in isolation as a technical issue, we can usually do a pretty good job of measurement. Subject experts are indeed the ultimate repository of validity. They accumulate a good deal of expertise in not only knowing what it is they want to measure, but what sort of questions produce pretty reliable answers. On a large scale, in cases where item banks of questions are a reasonable sampling mechanism, we know a lot – though we do not, in this country, always practise it – about how to develop reliable mass-market tests. But that is only a small part of measurement, and if we persist in concentrating only on the technical, we will continue to find ourselves in situations where expensive measurement systems conspicuously fail to measure what they were set up for.

If the answer to 'What can we measure?' is "Less than we think', then what next.?What, in educational assessment, should we be measuring?

Here I am going to start, rather than finish, with the answer. It is "One thing at a time.' Let me explain.

Educational assessment, in modern societies, serves a number of different functions.

Defining the curriculum and structuring teaching (and learning)

Sorting and selecting students

Certifying attainment

In addition formal assessment may be used, within the classroom, as a formative tool for helping individual pupils. It may also be used, at a system-wide level, as a way of ensuring accountability, and that government money is being used appropriately and

efficiently. The UK (especially England) has been a leader in this last area, with league tables and key stage tests. The latter, in particular, have proven of considerable interest to other countries. They are seen as a way of asserting control over school systems in situations where, in government eyes, there has been 'producer capture', with the teachers running the schools pretty much as they see fit.

Accountability systems are worth a whole lecture in themselves and I will just say two things here. First, I question both the effectiveness and the necessity of making them a bureaucratised central government function. If you think schools and universities need to be more 'accountable' then it is students and parents who need more information and power. And they can also be relied upon to read the labour market rather more accurately than meetings of senior businessmen summoned to Whitehall to opine on 'what business needs.' Second if you load an accountability function onto an assessment system which is already trying to do three very different things, you will increase, significantly, the likelihood of distortions, and the probability that none of those three prime functions will be carried out very well.

If we look at the three primary functions of educational assessment, we can see that, in 'pure' assessment terms they have very different implications – both for the style and approach they demand and in terms of how far classic reliability issues take centre stage.

In curriculum terms, there is general consensus that we are after a broad curriculum; that we want to be flexible in our capacity to incorporate new knowledge and insights, that we want there to be room for imaginative and inspiring teachers, room for students to be stretched and follow their interests, room for diversity. We also know that if parts of a curriculum are not assessed at all (and others are) everyone will either dump or devalue the un-assessed element. So what this calls out for is assessments which are not too hung up on reliability and uniformity, which are assessed using very broad categories, and where having completed something successfully is pretty much the sole requirement for success. (This is, of course, the pattern followed for large amounts of high-status professional training even if, these days, it is always dressed up in all sorts of language about outcomes.)

Sorting and selecting is a central function of assessment in any society and certainly in ours. (I hardly need point this out when standing in Cambridge.) It has fundamentally different features and requirements. You cannot – I repeat cannot – put people in a single rank-order which has any claims to objectivity, reliability ,or fairness by using complicated equating mechanisms to align completely different tests and measures. You cannot do it statistically and you cannot do it by black magic. If you want a fair rank order you need a single shared measure. (For those of you with an interest in this issue I refer you to an excellent article by Paul Newton in *Assessment in Education* 2005.) So sorting and selection requires common measures of the candidates – but also measures whose key characteristic is reliable reporting of fine grained differences in performance.

Thirdly, accreditation is different again. There was a fashion, not so long ago, for having 'criterion referenced' measurement for everything – which is fine until you start trying to decide whether someone who has achieved A, C and E but not B, D and F is better, worse or the same as someone in the mirror position. However, criterion-referencing of some sort is the sort of approach which is natural when you

are accrediting somebody. Driving tests are obvious examples but so are vocational and professional certificates. And some educational tests may also be best conceptualised in this way, especially some English and maths and language ones. What I hope is obvious is that the nature of the measurement, and the way it is designed, will again be different . What will be important is performance around a critical boundary which marks off performance which can be accredited from performance which falls short.

One of the major problems in this country has been, in my view, our addiction to using assessments for more than one purpose at once. This is particularly the case with GCSEs and A levels and the result is that each of the functions get carried out less well than it could be. One reason we do this may be an unconscious belief that it is more efficient this way and that we will not need to assess as much. However, since the response, when one overburdened test fails to do any of its jobs very well is usually to reform it in a way that increases the amount of assessment going on, it is time this unconscious conviction was challenged. .

To illustrate my point let me return to the conundrum with which I started. Mathematics in the sixth form is in a state of crisis, with too few students studying it. This is not because they have always hated it – at primary school it comes in third place after art and sports in children's preferences. It is not because they do not think it is important. The rewards of the financial sector, and the wages now going to engineering and other quantitative graduates are something more and more people are aware of, and many maths teaches by now know that Maths A level is going to increase your life time earnings.

A large part of the problem is certainly the teacher shortage. But this is itself part of a downward spiral that started some time ago, and is closely linked, in my view, to the way our assessment system works. A levels do not have much to do with certification but do have a lot to do with selection, and with curriculum definition. If a student wants to study maths at university, they will be selected from among a group who have done similar subjects with similar, rankable measurement data. However, most students who think of taking Maths A level are not would-be mathematics undergraduates. What they need is as many A grades as possible and to minimise the risk of any Fails. In a system which treats A levels as 'equivalent' this is a powerful disincentive.

The problem is exacerbated by pressure to steer students, at GCSE level, towards the syllabus on which they will most likely get a B (or be sure of a C.) In the year that a B grade first became available on the middle tier papers, the numbers of students taking the top tier papers halved, instantly. Yet at GCSE levels, maths examinations do serve a certification function – and research I have done suggests that a good deal of the vocal dissatisfaction about maths attainment expressed by engineering firms and within 'technical' FE is associated with the change in what GCSE grades 'mean' in terms of material covered by teir typical apprentice or entrant.

So the answer to the second question I posed is 'one thing at a time' – and in a manner appropriate to the task at hand. While on any given occasion, that may be complicated, or technically demanding, thinking this way also surely provides a better way of organising our assessment and measurement activities than does pursuit of perfect reliability and the 'true score'.