The meanings and consequences of educational assessments

Dylan Wiliam

Critical Quarterly, **42**(1), pp105-127 (2000)

Introduction and overview

The reasons for carrying out educational assessments can be grouped under three broad headings:

formative supporting learning
summative certifying individuals
evaluative holding educational institutions accountable

In this lecture, I want to argue that current policies with regard to the use of educational assessments, particularly those used in schools, have taken a wrong turn. They have started from the idea that the primary purpose of educational assessment is selecting and certifying the achievement of individuals (ie summative assessment)—and have tried to make assessments originally designed for this purpose also provide information with which educational institutions can be made accountable (evaluative assessment). Educational assessment has thus become divorced from learning, and the huge contribution that assessment can make to learning (ie formative assessment) has been largely lost. Furthermore, as a result of this separation, formal assessment has focused just on the outcomes of learning, and because of the limited amount of time that can be justified for assessments that do not contribute to learning, has assessed only a narrow part of those outcomes. The predictability of these assessments allows teachers and learners to focus on only what is assessed, and the high stakes attached to the results create an incentive to do so. This creates a vicious spiral in which only those aspects of learning that are easily measured are regarded as important, and even these narrow outcomes are not achieved as easily as they could be, or by as many learners, were assessment regarded as an integral part of teaching.

In place of this vicious spiral, I propose that developing a system that integrates summative and formative assessment will improve both the quality of learning and the quality of the assessment. A separate system, relying on 'light sampling' of the performance of schools would provide stable and robust information for the purposes of accountability and policy-formation.

I begin with a survey of current practices in national assessment, focusing in turn on the two key issues of reliability and validity.

Reliability

All measurements whether physical, psychological or educational, are unreliable. If we wanted to find out how accurate an instrument for measuring length was, one way to do that would be to take lots of measurements of the same length. These measurements would not all be the same , but we would expect them to cluster around a particular value, which we would regard as the best estimate of the actual length of whatever we were measuring. However, in doing this, we are assuming that the object we are measuring isn't changing length between our measurements. In other words, we *assume* that the differences between our measurements are 'errors of measurement'. This is a circular argument—we infer physical laws by ignoring variation in measurements that we assume to be irrelevant, *because of our physical law* (Kyburg, 1992). If our measurements cluster tightly together, we conclude that we have a reliable instrument, and if they are widely scattered, then we conclude that our instrument is unreliable.

In the same way, if we wanted to find out the reliability of an educational test, we would test a group of students with the same (or a similar) test many times. For each candidate, the marks obtained would cluster around a particular score, which we call the 'true score' for that candidate. This does not, of course, mean that we think that the individual being tested has a true ability or anything like that—the idea of a true score is just the long-run average of all the different scores if we test the same individual lots of times. This is the same as the circular argument we have to use in physical measurement to 'pull ourselves up by our own bootstraps'.

Now if the test is a good test, the values will cluster together tightly—the score that the candidate gets on any one occasion will be close, if not identical, to the score obtained on a second occasion. A bad test, on

the other hand will produce values that vary widely, so that, to a very real extent, the mark obtained on one occasion is not a reliable guide to what they would achieve on a subsequent occasion.

The reliability of a test is defined as a kind of 'signal to noise' ratio. The mark of any individual on a particular tests is assumed to be made up of a signal (the true score) and some noise (the error), and of course, we want the ratio of signal to noise to be as large as possible. This can be achieved in two ways. The best way would be to reduce the 'noise', by reducing the error in the test scores. However, we can also improve the signal to noise ratio by increasing the strength of the signal.

In the context of educational assessment, 'increasing the signal' entails making the differences between individuals as large as possible (in the same way that in communications engineering, say, increasing the signal would correspond to maximising the potential difference between presence and absence of signal in a wire). We want some students getting very high scores and others getting very low scores, so that the differences in test scores caused by the unreliability of the test are small compared to differences in the 'true' scores. This means that our test must *discriminate* between stronger and weaker candidates. To do this, we must select for our test items which are answered correctly only by those candidates who get high scores on the test overall. Other items, for which the success-rate of strong candidates is comparable to that for weaker candidates, are dropped because they do not discriminate. This has two consequences. The first is that choosing only what are regarded as 'good' items in this way invalidates all the statistical theories that are used to interpret test results, as was noted by Jane Loevinger over thirty years ago:

> Here is an enormous discrepancy. In one building are the test and subject matter experts doing the best they can to make the best possible tests, while in a building across the street, the psychometric theoreticians construct test theories on the assumption that items are chosen by random sampling. (Loevinger, 1965 p147)

The second consequence is that our attempts to make reliable tests guarantee the production of tests that maximise differences between individuals, and minimise differences between their experiences. This is why, for example, that many studies of school effectiveness have found that schools have comparatively little effect on educational achievement. In these studies, educational achievement has been assessed with a test that was designed to *maximise* differences between *individuals*, irrespective of the differences in their experiences. Such tests are bound to minimise the differences between schools.

The elevation of reliability as the primary criterion of quality is itself a value-laden assumption which has its roots in particular applications of educational assessment. As Cleo Cherryholmes remarks, "Constructs, measurements, discourses and practices are objects of history" (Cherryholmes, 1989 p115). The requirement for reliability, originally intended to ensure that the meaning of an assessment result was stable, turns out to *create*, and *reify* constructs—constructs that are generally assumed to be assessed in a value-free way.

The reliability of educational assessments

Educational assessments are unreliable for a number of reasons. The individuals being tested are not consistent in their performance—people have 'good days' and 'bad days'—and, apart from multiple-choice tests, there is also some inconsistency in the ways that assessments are marked. But the most significant cause of unreliability is the actual choice of items for a particular test. If we have an annual assessment like the national curriculum tests or GCSE examinations, then, because the papers are not kept secret, new versions have to be prepared each year. The question is then are the tests interchangeable? The two tests might be assessing broadly the same thing, but one of the two tests might suit a particular candidate better than the other. We therefore have a situation where the scores that candidates get depend on how lucky they are—reputedly the method that Napoleon used to choose his generals, and look where that got him!

In fact, as Robert Linn and his colleagues have shown, the unreliability caused by the variation in tasks is actually greater than that caused by disagreements amongst markers (Linn & Baker, 1996). What this means is that given a choice between a three-hour exam where each paper is marked by two markers, and six-hours of exams where each paper is marked only once, the latter produces greater reliability.

So, given these kinds of unreliability, how accurate are the results educational tests? The reliability of specialist psychological tests can be as high as 95%, while the reliability of examinations and tests used in education is typically of the order of 85%. Unfortunately, what this actually means in practice is not at all straightforward, so I will illustrate the consequences of the unreliability of typical educational tests.

An educational test will generally be pitched so as to have an average mark around 50% with the marks of the candidates ranging from about 20% to 80%. Let us consider what this means for a class of 30 children.

For half the children in the class, the mark that they get on any one occasion will be within 4% of their 'true score'—that is the long-run average of what they would get over many testing occasions. This is quite re-assuring, but the corollary of this is that for the other half of the students in the class, the mark they actually get is 'wrong' by more than 4 marks. And for one student in the class, the mark they get will be wrong by more than 12 marks. The student probably wouldn't mind if they get 12 marks *more* than they should have got, but it is equally likely that this error is the other way, and they get 12 marks less than they should have done. Of course, the student won't know this because they don't know what their true score is.

Now does this matter? Well, although a score 12 marks below what one should have got is very unfortunate for the individual concerned, if only one person in a class of thirty is seriously disadvantaged by an assessment process, we might consider this a price worth paying. However, it turns out that serious disadvantage in educational assessment is not that uncommon.

The government has never published reliability statistics for any of its statutory national curriculum assessments nor has it required the examination boards to publish statistics on the reliability of GCSE and A-level examinations. Indeed, one of the most remarkable features of the examination system of England and Wales is that relatively few reliability studies have been conducted. Those that have been carried out have found the reliability of educational assessments to be around 85%.

The fact that educational assessments are unreliable is accepted to an extent in this country. We are suspicious of percentage scores, and prefer, instead, to report grades in the case of school examinations and classes in the case of undergraduates degrees. And in a way this is very sensible, because although, on balance, it is likely that someone who got 65% on a test is likely to have a higher 'true score' than someone who got 64%, it could easily be the other way round, and even if the first person does have a higher true score, they are unlikely to be that much better. In response to the danger of claiming 'spurious precision' for scores, the tendency in the UK has therefore been to report not scores but grades (in school examinations) or classes (in university examinations). However, the result of reporting scores as grades is that it is too often assumed that the grades or classes are 'right'.

The grades or classes reported are likely to be 'right' when the score that an individual receives is right in the middle of the range for a particular grade or class, but when someone is close to the borderline between two grades, only a small error in their score will tip them over into a different grade. For example, suppose a university decides that candidates need a particular pattern of A-level grades—say three Bs—to benefit from a programme, then a student who gets two Bs and a C may well be rejected. If the cut-off for a grade B was 60%, and that for a grade A was 70%, then a student with marks of 60, 60 and 60 would get three Bs and would get in, but a student getting marks of 59, 69 and 69 would get two Bs and a C, and would probably be rejected. Had the admissions tutor been told the actual scores, it would be clear that the candidate had only just missed the threshold for B on the first subject, and given the other scores, may well have admitted her. If scores and percentages are prone to spurious precision, then grades are prone to spurious *accuracy*.

So, how accurate are reported grades? Well because of the lack of published statistics on educational assessments, we have to make some assumptions, but assuming that examinations have a reliability of 85%, and assuming that we use these examinations to allocate students to one of eight grades (as is the case in GCSE), then only about 60% of the candidates would get the 'right' grade. Of the remaining 40%, half would get a higher grade than they should, and half would get a lower grade than they should, and for a small number of candidates, their reported grade will be out by *two* grades. For school tests and examinations, the government's response to this has been to abandon the attempt to distinguish between eight different levels of performance in one examination, and instead have different examinations for students of different levels of achievement. In mathematics at GCSE, for example, there are three tiers of examination each of which gives access to only four or five grades. With tiering, the number of candidates getting the 'wrong' grade will be reduced, but only at the cost of restricting the grades available to candidates. For example, candidates who take the least demanding tier cannot get a good grade (ie grade 'C' or higher) no matter how well they do. This places a great deal of pressure on teachers to make the right entry choice, and produces considerable alienation amongst the students whose potential achievement is restricted in this way.

The important point here is that these difficulties arise because of a fundamental limitation in the accuracy we can expect of traditional timed examinations. The introduction of 'tiering' does increase the proportion of candidates correctly classified, but only at the cost of mis-classifying others, and more importantly, alienating a far greater number of students, who, because of this alienation may well give up and fail to achieve the grades of which they are capable—a cure that is probably worse than the disease.

This debate about the costs and benefits of tiering has been conducted largely within the professional and academic community. Given the importance attached by the public to these results, the absence of any public concern about the lack of information about the reliability of educational assessments is rather puzzling. In this country, opinion pollsters routinely publish margins of error for their poll results, and in the United States, it is expected that any user of test information will know the limits of the test result they are using. And yet, in this country, we treat test results as perfectly accurate. Why is there no measurement error in the UK?

One perspective on this is provided by the work of J L Austen who in the 1955 William James lectures, discussed two different kinds of 'speech acts'—illocutionary and perlocutionary (Austin, 1962). Illocutionary speech acts are those that by their mere utterance actually do what they say. In contrast, perlocutionary speech acts are speech acts *about* what has been, is or will be. For example, the verdict of a jury in a trial is an illocutionary speech act—it does what it says, since the defendant becomes innocent or guilty simply by virtue of the announcement of the verdict. Once a jury has declared someone guilty, they *are* guilty, whether or not they really committed the act of which they are accused, until that verdict is set aside by another (illocutionary) speech act. Another example of an illocutionary speech act is the wedding ceremony, where the speech act of one person (the person conducting the ceremony saying "I now pronounce you husband and wife") actually does what it says, creating the 'social fact' of the marriage (Searle, 1995).

The idea of the accuracy of an assessment derives from a view of assessment results as *perlocutionary* speech acts. If we claim to be describing someone's performance now, in the past, or predicting their performance in the future, then it makes sense to ask how accurate that description is, which often raises questions of objectivity and subjectivity. However, while it may make sense to question the *authority* of a maker of a speech act to create social facts, it does not make any sense to question the *accuracy* of those speech acts. This point is well illustrated by the story of the journalist asking an American baseball umpire whether his judgements were subjective or objective:

Interviewer: Did you call them the way you saw them, or did you call them the way they were?

Umpire: The way I called them *was* the way they were.

Rightly or wrongly, in the United Kingdom, at the moment, the pronouncements of the government's testing agencies are treated as illocutionary speech acts, creating the social fact of an individual's success or failure. The grade you get is the grade you get, and arguing about the likely effect of measurement error will do you no more good than claiming to an umpire that you weren't out—he'll just tell you to look in the newspaper tomorrow...

Such an authoritarian stance is tenable in a stable social order, but at a time when the authority of professionals is (in my view rightly) open to challenge, it is a dangerous tactic. The reliability of our national assessments is simply not good enough to warrant the trust that is placed in them. And one day, people are going to find this out.

What is perhaps even more surprising, is that reliability (ie how accurately we are measuring something) has been a priority in the development of our national assessment systems and has arguably taken precedence over questions of validity (ie whether we a re measuring the right thing). This is perhaps best summed up by the story of the drunk looking for his keys at night under a streetlamp. When asked, "Is this where you dropped them?", he replies, "No, but this is where the light is".

Validity

Instead of asking how accurately are we measuring something, a concern for validity asks what, exactly are we measuring—specifically what do the results of our assessments *mean*?

The traditional definition of validity—and one that dates back at least sixty years—is that an assessment is valid if it assesses what it purports to assess. This definition is still common in many texts on assessment in this country even though it is unsatisfactory in many respects.

In the first place, validity cannot be a property of an assessment. If we have a science test that happens to be written in a way that requires a high level of reading skill to discover what the questions are asking, then this test may well be a good science test for fluent readers, but it will not be a good test for poor readers. In other words, the validity of a test can change according to who takes the test.

In the second place, an assessment does not purport anything—it tests simply what it tests. The purporting is done by those who claim that a particular test result tells us something beyond just the result of that test. This is the fundamental issue in educational assessment—how we can move from a candidate's score on a particular assessment to making more general claims. This is why it has become increasingly accepted over the last thirty years that validity is not a property of a test at all, but a property of the conclusions that we draw on the basis of test results.

This marks a huge shift, because it transfers some, if not the majority, of the responsibility for establishing validity from those who make tests to those who draw specific conclusions about the meaning of test results. In the words of Lee Cronbach, "One validates, not a test, but an *interpretation of data arising from a specified procedure*"(Cronbach, 1971 p447, emphasis in original).

For example, with the traditional view of validity, the responsibility for the validation of A-level examinations would fall on the examination boards. It would be up to them to show that the exams did actually assess 'Physics' or 'English Literature'. This would involve showing that the examinations did assess the syllabuses published by the examination board.

However, we also use examination results in other ways. Rather than interpreting examination results to tell us how well students have done in the past, we also use them to attempt to predict how well they will do in the future.

Universities want to select students who will do well at university, and of course we can't really find this out until the students have actually been to university. What we can do, however, is to look for something that correlates well with the outcomes of university education, but which can be assessed before students go to university. For most students, the measure that is used is A-level, but with this new conception of validity, universities who wish to use A-level scores for deciding which students they admit must provide evidence that the use of A-levels in this way is warranted.

Within this view, validity subsumes reliability, because any conclusions we might want to draw from a set of test results are unlikely to be justifiable if the same test administered to the same candidate could generate completely different result tomorrow.

The use of tests to predict future performance is quite widespread. In some local authorities, scores of 11+ tests are used to predict the capability of learners to "benefit from a grammar school education", GCSE scores are used to select which students should go on to do A-level and A-level scores are used to predict who should go on to university. The extent to which an assessment can be used to predict future performance is usually expressed by a correlation—a good predictor is one where students getting high scores on the predictor go on to get good scores at the next level. Again, studies of how good these predictors are few and far between, but correlations around 70% are typical. What this means in practice depends how selective we are being, but if we are selecting around one-third of the individuals applying, then with a typical selection test, we would only be making the right decision for around three-quarters of the people. For the other quarter of the population, we would either be taking those who we shouldn't, or not taking those who we should. Given the inaccuracy of this selection process, we ought to think very hard about whether we need to select at all.

The same issues apply to the idea of 'targeting' particular students, which has become very popular in many secondary schools in England and Wales. Because the results of IQ tests taken at the age of 11 are correlated with GCSE scores at the age of 16, schools believe they can identify the particular students who they can expect to achieve the government's key performance indicator for secondary schools—five 'good' grades at GCSE. When a large proportion of these 'targeted' students achieve the grades expected, the school believes that its targeting is working. However, since the correlation between IQ at 11 and GCSE scores at age 16 is only around 70%, a large number of students who could have achieved those grades, had they received the extra support given to 'targeted' students, do not. When predictor and outcome variables have a correlation of only 70%, the important point is that it is still all to play for. The technology of selection and targeting is not reliable or valid enough for the claims that are made for them. We simply do not know who has the potential to do well, and therefore, what we need are inclusive systems of education that allow *all* students to be targeted.

Now the analysis so far has depended on traditional ideas of reliability and validity—ones that are, to all intents and purposes, directly borrowed from mainstream psychological testing. With educational assessments, however, we cannot ignore the fact that these assessments are carried out by, on and for real people, who change what they do as a result of the assessment. It is for this reason that Samuel Messick proposed in 1980 that a consideration of the *consequences* of the use of educational assessments should be part of the process of validation.

For example, most science teachers agree that practical skills are an important part of the content of a science curriculum. An assessment of 'science' therefore, ought to assess practical skills as well as more traditional forms of scientific knowledge and capability. However, testing practical skills is expensive, and those concerned with the efficiency of the assessments point out that the results of the practical and written tests correlate very highly, so there's no need to carry on with the expensive practical testing. The same sorts of arguments have dominated the debate in the United States between multiple-choice and constructed-response tests. What then happens is the practical aspects of science are dropped from the assessment. The consequence of this, for the domain of school science is to send the message that practical science isn't as important as the written aspects. The social consequence of this is that teachers, understandably anxious to get their students the best possible results in the assessment, not least because of its influence over the students' future career prospects, place less emphasis on the practical aspects of science. Because teachers are no longer teaching practical science hand-in-hand with other aspects of science, the correlation between students' performance in practical aspects of science and the written aspects weakens, so that it is no longer possible to tell anything about a student's practical competence from the score on the science assessment. This is an example of what has become known as Goodhart's law, name after Charles Goodhart, a former chief economist at the Bank of England, who showed that performance indicators lose their usefulness when used as objects of policy.

The example he used was that of the relationship between inflation and money supply. Economists had noticed that increases in the rate of inflation seemed to coincide with increases in money supply, although neither had any discernible relationship with the growth of the economy. Since no-one knew how to control inflation, controlling money supply seemed to offer a useful policy tool for controlling inflation, without any adverse effect on growth. And the result was the biggest slump in the economy since the 1930s. As Peter Kellner comments, "The very act of making money supply the main policy target changed the relationship between money supply and the rest of the economy" (Kellner, 1997).

Similar problems have beset attempts to provide performance indicators in the Health Service, in the privatised railway companies and a host of other public services. A variety of indicators is selected for their ability to represent the quality of the service, but when used as the sole index of quality, the manipulability of these indicators destroys the relationship between the indicator and the indicated.

A particularly striking example of this is provided by one state in the US, which found that after steady year-on-year rises in state-wide test scores, the gains began to level off. They changed the test they used, and found that, while scores were low to begin with, subsequent years showed substantial and steady rises. However, when, five years later, they administered the original test, performance was way below the levels that had been reached by their predecessors five years earlier. By directing attention more and more onto particular indicators of performance they had managed to increase the scores on the indicator, but the score on the indicated was relatively unaffected.

Now in the past, I have argued that if schools are to be held accountable via measures of performance, then the 'value-added' by the school—that is the amount of progress made by a student at the school—is a better measure than the achievement of students on leaving, which as often as not tells us more about what they knew when they started at the school. But the manipulable nature of the assessments that we are using means that both raw scores and value-added analyses are likely to be almost meaningless. The government is already finding that while it is making steady progress towards its targets for 11-year-olds, this has not been matched by progress towards its targets for 14-year-olds. Now part of this is no doubt due to the fact that extra money has been provided for 'catch-up' classes for 11-year-olds, but the reason this has helped is really only because the tests for 11-year-olds are easier to coach students for than those at age 14, combined with the fact that many secondary schools see the tests for 14-year-olds as irrelevant compared to the GCSE.

Some authors have argued that the social consequences of test use, although important, are beyond the concerns of validity as such. However, others, notably the late Samuel Messick, have argued that where the use of assessments changes what people do, any enquiry into the quality of assessments that ignores the social consequences is impoverished. He has proposed that an argument about the validity of an assessment, for a particular use, requires the simultaneous consideration of four strands of evidence:

a evidence that the scores have a plausible meaning in terms of the domain being assessed
b evidence that the predictions that will be made from the results are justifiable
c an evaluation of the value implications inherent in adopting the assessment
d an evaluation of the social consequence of using the assessment

These four strands can be presented as the result of crossing two facets as shown in figure 1. To illustrate this it is instructive to consider the argument that took place in 1991 about the relevant weighting of the three assessment components that were to be combined in order to produce an overall level for a student's

achievement in English at age 14. The test developers had proposed that the three components (Speaking and Listening, Reading, and Writing) should be equally weighted, while the government's advisers wanted to use the ratio 30:35:35. Within the classical validity framework, this would appear to be a technical debate about which weighting scheme would provide the best description of a candidate's performance in English, or which one best predicted future success in the subject. In fact, on a sample of 2000 students for whom the component scores were available, only two changed level from one weighting scheme to the other! The heat of the debate that was occasioned by this issue is therefore hard to understand. However, within Messick's framework, we can see that while there is little to choose between the two weighing schemes in terms of the *meanings* of the results, they differ markedly in their *consequences*. Giving less weight in the mark scheme to Speaking and Listening than Reading or Writing sends the message that Speaking and Listening is less important than Reading and Writing. In other words, control of the mark scheme allows one to send messages about the values associated with an assessment. The presumed social consequence of this is to persuade teachers then to place greater emphasis (and therefore, presumably, more teaching time) on Reading and Writing than Speaking and Listing,

	within-domain	beyond-domain
meanings	construct validity (content considerations)	construct validity (predictive and concurrent validity)
consequences	value implications	social consequences

Figure 1: facets of validity argument (after Messick, 1980)

Messick's model provides an integrated view of validity as "an overall evaluative judgement of the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1988 p42), which takes into account the essentially social nature of educational assessments. Put simply, a test is valid to the extent that you are happy for teachers to teach towards the test, because, were this the case:

- the only way to increase the student's score on the test would be to increase the score on the whole of whatever the test is meant to be testing.
- increasing the student's score on the test would indicate an improvement in whatever the test is used to predict.
- the value implications of the test—ie what messages it sends about what is important—would be appropriate; and
- the social consequences of the likely uses of the test would be acceptable.

To sum up, the trouble with the prevalent approach to educational assessment in this country is that we have divorced the certification of achievement and capability from the learning process. Because the assessments we use have no educational value, we feel unable to justify spending a lot of time on them, so that we typically assess the outcome of several thousand hours of learning with assessments that last only a few hours. Giving so little time to the assessment means that we can assess only a limited proportion of what has been taught, and conducting the assessments in 'standardised conditions' means that teachers can easily guess which parts of the curriculum are going to be assessed. Because of the importance attached to these outcomes, teachers and students are pressured into focusing on only those aspects of the curriculum that will be assessed.

This process is well summed up by Charles Handy's rendering of the what has come to be known as the Macnamara Fallacy, named after the US Secretary of Defense:

> The Macnamara Fallacy: The first step is to measure whatever can be easily measured. This is OK as far as it goes. The second step is to disregard that which can't easily be measured or to give it an arbitrary quantitative value. This is artificial and misleading. The third step is to presume that what can't be measured easily really isn't important. This is blindness. The fourth step is to say that what can't be easily measured really doesn't exist. This is suicide. (Handy, 1994 p219)

We started out with the aim of making the important measurable, and ended up making only the measurable important.

A different starting point

Over the last ten years, there have been three major reviews of the contribution that informal classroom assessment, conducted as part of teachers' day-to-day activities, can make to raising standards of achievement (Black & Wiliam, 1998; Crooks, 1988; Natriello, 1987). The message from these studies, between them covering over 500 research studies, is clear. Improving the quality of teachers' day-to-day assessment practices has a substantial effect on the achievement of students—big enough to take an average country in the international 'league tables' of student achievement, such New Zealand, England or the United States, up into the top 5. The essential feature of effective classroom assessment is not merely that it is diagnostic (ie tells learners where they are going wrong) but also formative (ie tells them what to do in order to improve).

Strictly speaking, of course, there is no such thing as a formative assessment. The distinction between formative and summative applies not to the assessment itself, but to the use to which the information arising from the assessment is put. The same assessment can serve both formative and summative functions, although in general, the assessment will have been designed so as to emphasise one of the functions. The defining feature of a formative assessment is that *the information fed back to the learner must be used by the learner in improving performance*. If, for example, the teacher gives feedback to the student indicating what needs to be done next, that is not formative unless the learner can understand *and act* on that information. An essential pre-requisite for assessment to serve a formative function is therefore that the learner comes to understand the goals towards which she is aiming (Sadler, 1989). If the teacher tells the student that she needs to "be more systematic" in her mathematical investigations, that is not feedback unless the learner understands what "being systematic" means—otherwise this is no more helpful than telling an unsuccessful comedian to "be funnier". The difficulty with this is that if the learner understood what "being systematic" meant, she would probably have been able to be more systematic in the first place. The teacher believes the advice she is giving is helpful, but that is because the teacher already knows what it means to be systematic.

Of course, to be practicable, such assessments must be built into the teacher's day-to-day practice, but once achieved, a formative assessment system would generate a wealth of data on the achievements of the individual student. Some of these—particularly where a teacher has probed a particular aspect very deeply—would be more use in determining a student's future learning needs than in establishing an overall level of achievement, so it would be unwise to use any mechanistic formula in order to move from the fine-grained record to an overall level of achievement for the year. Instead, the teachers would re-interpret all the available data in order to come up with a grade for the student.

The immediate reaction of many to this proposal is that such a system cannot be used for high-stakes assessment, such as school-leaving and university entrance examinations because teachers' judgements of their students cannot be objective. In a sense this is true. Teachers' judgements of their own students will not be objective, but then on the other hand neither will any other kind of assessment.

Any assessment just tells us what the student achieved on that assessment. In this sense, a test tests only what a test tests. However, we are hardly ever interested in the result of the test *per se*. We are generally interested in the results of a test as a sample of something wider, and so we interpret each test result *in terms of something else*.

Norm-referenced and cohort-referenced assessments

For most of the history of educational testing, the way we have made sense of test results has been by comparing the performance of an individual with that of a group of students, and this has been done in one of two ways. The first is to compare the performance of an individual with the others who took the test at the same time. This is often called a norm-referenced test (more precisely a norm-referenced interpretation of a test result), but in fact it is better termed a *cohort*-referenced test, since the only comparison is with the cohort of students who took the test at the same time. If we want to select thirty people for a particular university course, and we have a test that correlates highly with the outcomes of the course, then it might be appropriate to admit just the thirty students who get the highest score on the test. In this case, each candidate is compared only with the other candidates taking the test at the same time, so that sabotaging someone else's chances improves your own. Such a test is truly competitive.

Frequently, however, the inferences that are sought are not restricted to just a single cohort and it becomes necessary to compare the performance of candidates in a given year with those who took the same assessment previously. The standard way to do this is to compare every candidate that takes a test with the performance of some well-defined group of individuals. For example until recently, the performance of

every single student who took the American Scholastic Aptitude Test (SAT) was compared with a group of college-bound young men from the east coast of the United States who took the test in 1941.

Both norm- and cohort-referenced assessment are akin to the process of 'benchmarking' in business, which is tantamount to saying "we have no idea what level of performance we need here, so let's just see how we're doing compared with everybody else". All that is required for this is that you can put the candidates in rank order. The trouble with this approach is that you can very easily put people in rank order without having any idea of what you are putting them in rank order *of*.

It was this desire for some clarity about what actually was being assessed, particularly for teaching purposes, that led to the development of criterion-referenced assessment in the 1960s and 1970s.

Criterion-referenced assessments

The essence of criterion-referenced assessment is that the domain to which inferences are to be made is specified with great precision. In particular, it was hoped that performance domains could be specified so precisely that items for assessing the domain could be generated automatically and uncontroversially (Popham, 1980).

However, as Angoff (1974) has pointed out, any criterion-referenced assessment is underpinned by a set of norm-referenced assumptions, because the assessments are used in social settings. In measurement terms, the criterion 'can high jump two metres' is no more interesting than 'can high jump ten metres' or 'can high jump one metre'. It is only by reference to a particular population (in this case human beings), that the first has some interest, while the latter two do not.

The need for interpretation of criteria is clearly illustrated in the UK car driving test, which requires, among other things, that the driver "Can cause the car to face in the opposite direction by means of the forward and reverse gears". This is commonly referred to as the 'three-point-turn', but it is also likely that a five point-turn would be acceptable. Even a seven-point turn might well be regarded as acceptable, but only if the road in which the turn was attempted were quite narrow. A forty-three point turn, while clearly satisfying the literal requirements of the criterion, would almost certainly not be regarded as acceptable. The criterion is there to distinguish between acceptable and unacceptable levels of performance, and we therefore have to use norms, however implicitly, to determine appropriate interpretations.

Another competence required by the driving test is that the candidate can reverse the car around a corner without mounting the curb, nor moving too far into the road, but how far is too far?' In practice, the criterion is interpreted with respect to the target population; a tolerance of six inches would result in nobody passing the test, and a tolerance of six feet would result in almost everybody succeeding, thus robbing the criterion of its power to discriminate between acceptable and unacceptable levels of performance.

In any particular usage, a criterion is interpreted with respect to a target population, and this interpretation relies on the exercise of judgement that is beyond the criterion itself. In particular, it is a fundamental error to imagine that the words laid down in the criterion will be interpreted by novices in the same way as they are interpreted by experts. For example, the national curriculum for English in England and Wales specifies that average 14-year olds should be able to show "sensitivity to others" in discussion. The way that this is presented suggests that "sensitivity to others" is a prior condition for competence, but in reality, it is a *post hoc* description of competent behaviour. If a student does not already understand what kind of behaviour is required in group discussions, it is highly unlikely that being told to be 'sensitive to others' will help. What are generally described as 'criteria' are therefore not criteria at all, since they have no objective meaning independent of the context in which they are used. This point was recognised forty years ago by Michael Polanyi who suggested that intellectual abstractions about quality were better described as 'maxims':

> "Maxims cannot be understood, still less applied by anyone not already possessing a good practical knowledge of the art. They derive their interest from our appreciation of the art *and cannot themselves either replace or establish that appreciation"* (Polanyi, 1958 p50).

The same points have been made by Robert Pirsig who also argues that such maxims are *post hoc descriptions* of quality rather than constituents of it:

> Quality doesn't have to be defined. You understand it without definition. Quality is a direct experience independent of and prior to intellectual abstractions (Pirsig, 1991 p64).

How we make sense of assessment results therefore in most cases depends neither on a comparison with a norm-group, nor on the existence of unambiguous criteria. Instead, for the vast majority of assessments, what appears to be going on is that an individual result is interpreted in terms of the collective judgement of a community of examiners. I want to illustrate this by describing the practices of teachers who have been involved in 'high-stakes' assessment of English Language for the national school-leaving examination in England and Wales.

Construct-referenced assessment

In this innovative system, students developed portfolios of their work which were assessed by their teachers. In order to safeguard standards, teachers were trained to use the appropriate standards for marking by the use of 'agreement trials'. Typically, a teacher is given a piece of work to assess and when she has made an assessment, feedback is given by an 'expert' as to whether the assessment agrees with the expert assessment. The process of marking different pieces of work continues until the teacher demonstrates that she has converged on the correct marking standard, at which point she is 'accredited' as an assessor for some fixed period of time. However, even though these teachers have shown that they can apply the required standard of assessment consistently, their marking will still be double-checked by an external assessor, who has the power to amend the grades being awarded.

The innovative feature of such assessment is that no attempt is made to prescribe learning outcomes. In that it is defined at all, it is defined simply as the consensus of the teachers making the assessments. The assessment is not objective, in the sense that there are no objective criteria for a student to satisfy, but the experience in England is that it can be made reliable. To put it crudely, it is not necessary for the examiners to know what they are doing, only that they do it right.

In looking at what is really going on when people arrive at consensual judgements, Tom Christie and Gerry Forrest (Christie & Forrest, 1981) argued that the judgements of examiners was *limen*-referenced, suggesting that the examiners had a notion of a threshold standard that was required to receive a particular grade. Subsequently, Royce Sadler suggested that these judgements were *standards*-referenced (Sadler, 1987), indicating that they were arrived at by the assessors sharing a common standard for assessment. Both these ideas capture aspects of the process. The notion of a threshold is very familiar to experienced examiners who know that the crucial distinctions need to be made at the borderlines. However, there are two ways in which assessors might come to the judgment. The first is through the examiners having in their minds a clear notion of the relevant threshold. The second is that the assessors may come to their judgement not by looking at thresholds, but by having an idea of (say) a 'typical' D and a 'typical' C and seeing which is closer to the piece of work being assessed. The other difficulty with the idea of the notions of thresholds and standards is that they appear to suggest that the 'standards' in question lie along a single scale. This may well be true for simple assessments, but for the kinds of assessments that are used in 'high-stakes' assessments, there are many different routes to the same grade. In this sense, a particular grade level in an assessment appears to be more like the idea of a syndrome in medicine. In medicine, a syndrome is a collection of symptoms that often occur together. Generally, a patient exhibiting one or two of the signs associated with the syndrome would not be regarded as having the syndrome but with three or four, might be. However, it is possible that some of the characteristics may, in a particular individual, be strong enough for the syndrome to be established on the basis of one or two symptoms. In the same way whether a piece of work is a grade C or a grade D at A-level, or whether a thesis does or does not merit a PhD, involves balancing many different factors, but the important point is that the absence of some, or even the majority of the relevant factors does not mean that the piece of work is not worth the award being considered.

In order to encompass all these ideas, I have proposed that these assessments are in fact 'construct-referenced' because they rely on the existence of a shared construct of quality—shared between the community of assessors. The touchstone for distinguishing between criterion- and construct-referenced assessment is the relationship between the written descriptions (if they exist at all) and the domains. Where written statements collectively *define* the level of performance required (or more precisely where they define the justifiable inferences), then the assessment is criterion-referenced. However, where such statements merely *exemplify* the kinds of inferences that are warranted, then the assessment is, to an extent at least, construct-referenced.

The big question about any such system is of course, how reliable it is. Although a few studies of the reliability of teachers' assessment of English portfolios have been conducted, none has been published. One such study found that the teachers' grades agreed with the grades given by experts for 70% of the portfolios examined. This was considered unacceptably low, and so the results weren't published. However, the figure of 70% was the consistency with which students were given the correct grade, which as we have seen, is very different from traditional measures of reliability. Table 1 shows how the

proportion of students who would be correctly classified on an eight-grade scale, such as that used in GCSE, varies with the reliability of the marking.

Reliability	.60	.70	.80	.90	.95	.99
Grading accuracy	40%	45%	52%	65%	75%	90%

Table 1: impact of reliability of marking on accuracy of grading for an 8-grade scale

As can be seen, a grading accuracy of 70% is achieved only when the reliability of the marking is well over 0.90—a figure that has never been achieved in GCSE. Teachers' own assessments of their students' portfolios are, it seems more reliable than traditional timed written examinations.

The other concern that is raised by the replacement of examinations with coursework is that of authentication. Without formal written examinations, there is always a question mark over who's work is being assessed. In GCSE, for example, despite its name, coursework is done largely outside lesson time, so that there are real concerns that what is really being assessed is the access to learning resources such as encyclopaedias and computers at home, or even what the student's parents know. However, the consigning of coursework to time outside lessons is symptomatic of a view of assessment as divorced from day-to-day classroom work. In an integrated assessment system, there would be no question about the authentication of the work, because it would have been done by the students *in the class*. Coursework would be *course*work—the vehicle for learning rather than an addition to the load.

Accountability

The system I have outlined above would, I believe, allow the benefits of formative assessment to be achieved, while producing robust assessments of the achievements and capabilities of students. However, it does not address the issue of the narrowing of the curriculum caused by 'teaching to the test'. Even coursework-based assessments will be compromised if the results of individual students are used for the purpose of holding educational institutions accountable. To avoid this, if a measure of the effectiveness of schools is wanted, it could be provided by using a large number of tasks that cover the entire curriculum, with each student randomly assigned to take one of these tasks. The task would not provide an accurate measure of that student's achievement, because the score achieved would depend on whether the particular task suited the individual. But for every student at a school who was lucky in the task they were assigned, there would be one who was unlucky, and the average achievement across the tasks would be a very reliable measure of the average achievement in the school. Furthermore, the breadth of the tasks would mean that it would be impossible to teach towards the test. Or more precisely, the only effective way to teach towards the test would be to raise the standard of all the students on all the tasks, which, provided the tasks are a broad representation of the desired curriculum, would be exactly what was wanted. The government and policy makers would have undistorted information about the real levels of achievement in our schools, and users of assessment results would have accurate indications of the real levels of achievement of individual students, to guide decisions about future education and employment.

Summary

Current policies on testing and assessment start from the idea that the main purpose of educational assessment is selecting and certifying the achievement of individuals at the end of the 'key stages' of schooling (ie at ages 7, 11, 14, 16 and 18), and then have tried to make these tests also provide information to parents about the quality of schools. Because the tests in use have little educational value, they have to be short, and thus are unreliable, and test only a narrow range of the skills needed for life in the 21st century. Furthermore, because the tests are narrow, and test only what can easily be tested, schools have found it possible to guess what topics are going to come up in the tests, and can increase their test results by ignoring important topics that do not get tested. Focusing on the test results has meant that the contribution that day-to-day assessments can make to learning has been ignored.

However, research collected all over the world shows that if schools used assessment during teaching, to find out what students have learned, and what they need to do next, on a daily basis, the achievement of British students would be in the top five in the world, after Signapore, Japan, Taiwan and South Korea.

We have created a vicious spiral in which only those aspects of learning that are easily measured are regarded as important, and even these narrow outcomes are not achieved as easily as they could be, or by as many learners, if assessment was regarded as an integral part of teaching.

In place of this vicious spiral, I have proposed that the test results for individual students should be derived from teacher assessments, rigorously moderated by external assessors. A separate system, relying on ‘light sampling’ of the performance of schools would provide stable and robust information for the purposes of accountability and policy-formation.

Concluding note

There is a widespread view that the purpose of educational research is to lead practice—that educational researchers should experiment to find what works best, and encourage teachers to take this up. In my own work, however, I am acutely aware that I have not been leading but following behind the work of teachers and other assessors, trying to make sense of what they do, and, I hope, providing a language with which we can talk about what is going on with a view to sharing it more widely. I think it is appropriate, therefore, to conclude this lecture with the words of a teacher:

> [illegible] —Peter Silcock

References

Angoff, W. H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review,* **92**(Summer), 2-5, 21.

Austin, J. L. (1962). *How to do things with words : the William James Lectures delivered at Harvard University in 1955*. Oxford, UK: Clarendon Press.

Black, P. J. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles Policy and Practice,* **5**(1), 7-73.

Cherryholmes, C. H. (1989). *Power and criticism: poststructural investigations in education*. New York, NY: Teachers College Press.

Christie, T. & Forrest, G. M. (1981). *Defining public examination standards*. London, UK: Macmillan.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.) *Educational measurement* (pp. 443-507). Washington DC: American Council on Education.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research,* **58**(4), 438-481.

Handy, C. (1994). *The empty raincoat*. London, UK: Hutchinson.

Kellner, P. (1997). Hit-and-miss affair. *Times Education Supplement*, 23.

Kyburg, H. E. (1992). Measuring errors of measurement. In C. W. Savage & P. Ehrlich (Eds.), *Philosophical and foundational issues in measurement theory* (pp. 75-91). Hillsdale, NJ: Lawrence Erlbaum Associates.

Linn, R. L. & Baker, E. L. (1996). Can performance-based student assessment be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based assessment—challenges and possibilities: 95th yearbook of the National Society for the Study of Education part 1* (pp. 84-103). Chicago, IL: National Society for the Study of Education.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review,* **72**(2), 143-155.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist,* **35**(11), 1012-1027.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist,* **22**(2), 155-175.

Pirsig, R. M. (1991). *Lila: an inquiry into morals*. New York, NY: Bantam.

Polanyi, M. (1958). *Personal knowledge*. London, UK: Routledge & Kegan Paul.

Popham, W. J. (1980). Domain specification strategies. In R. A. Berk (Ed.) *Criterion-referenced measurement: the state of the art* (pp. 15-31). Baltimore, MD: Johns Hopkins University Press.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education,* **13**, 191-209.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science,* **18**, 145-165.

Searle, J. R. (1995). *The construction of social reality*. London, UK: Allen Lane, The Penguin Press.