

Integrating formative and summative functions of assessment¹

Dylan Wiliam²

King's College London

Introduction

In 1988, the British government began the phased introduction of a national curriculum accompanied by a system of national tests administered to all school students at the ages of 7, 11 and 14, with formal school-leaving examinations at the age of 16 (see Daugherty, 1995 and Wiliam, 1995a for detailed accounts). The effect of this programme of national testing has been a powerful demonstration of the truth of Goodhart's law.

This law was named 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 commented, "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 Britain's National Health Service, in the privatised railway companies and a host of other public services. Indicators are selected initially for their ability to represent the quality of the service, but when they are used as the main indices of quality, the *manipulability* (Wiliam, 1995b) 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 initially low, subsequent years showed substantial and steady rises. However, when, five years later, they administered the original test, performance was well 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 what these scores *indicated* was relatively unaffected (Linn, 1994). In simple terms, the clearer you are about what you want, the more likely you are to get it, but the less likely it is to mean anything.

What we have seen in England over the past ten years is a vivid demonstration of this. The raw results obtained by both primary and secondary schools in national tests and examinations are published in national and local newspapers. This process of 'naming and shaming' was intended by the government to spur schools into improvement. However, it turned out that parents were far more sophisticated in their analysis of school results than the government had imagined, and used a range of other factors apart simply from raw examination results in choosing schools for their children (Gewirtz, Ball, & Bowe, 1995). In order to increase the incentives for improvement even further, therefore, the government instituted a series of inspections with draconian powers (it is actually a criminal offence in England to deny an inspector access to data in a school). Schools have been inspected on a four-year cycle, but when the results obtained by a school are low—even though the attainment of the students at the school might have been well below the national average

¹ Paper presented to Working Group 10 of the International Congress on Mathematics Education, Makuhari, Tokyo, August 2000.

² Address for correspondence: Dylan Wiliam, School of Education, King's College London, Waterloo Road, London SE1 8WA. Tel: +44 20 7848 3153; Fax: +44 20 7848 3182; Email: dylan.wiliam@kcl.ac.uk

when they started at the school—government inspectors are sent in to the school outside the four-year-cycle. If they find the quality of teaching and learning at the school unsatisfactory, they return every month, and if no improvements are made, the school can be closed or ‘reconstituted’, and all the teachers can lose their jobs.

This creates a huge incentive for teachers to improve their students’ test and examination results at any cost. Even in primary schools, for up to six months before the tests, the teachers concentrate almost exclusively on the three subjects tested (English, mathematics and science), and a large number of ten and eleven-year-old students are suffering extreme stress (Reay & Wiliam, 1999). In secondary schools, because the primary measure of success focuses on a particular level of achievement (proportion of students achieving one of the four upper grades in at least five subjects) students close to the threshold are provided with extra teaching. Those considered too far below the threshold to have any reasonable chance of reaching it, on the other hand, are, if not ignored, typically taught by less qualified and less skilled teachers (Boaler, Wiliam, & Brown, 2000).

This has come about because we started from the idea that the primary purpose of educational assessment was selecting and certifying the achievement of individuals and then tried to make assessments originally designed for this purpose also provide information with which educational institutions can be made accountable. Educational assessment has thus become divorced from learning, and the huge contribution that assessment can make to learning (see for example the reviews of research in this area undertaken by Crooks, 1988; Natriello, 1987; and Black & Wiliam, 1998) 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, this formal assessment 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 to be regarded as an integral part of teaching. The root of this problem, I believe, is the way that we have conceptualised the distinction between formative and summative assessment.

Formative and summative functions of assessment

The assessment of educational attainment serves a variety of functions. At one extreme, assessment is used to monitor national standards. This is typically undertaken either to provide evidence about trends over time within a country—such as the National Assessment of Educational Progress programme in the United States or the Assessment of Performance Unit in England and Wales—or to compare standards of achievement with those in other countries (see Goldstein, 1996, for a brief review of the large-scale international comparisons carried out over the past 40 years). Educational assessments are also used to provide information with which teachers, educational administrators and politicians can be held accountable to the wider public. For individual students, educational assessments provide an apparently fair method for sorting and classifying students, thus serving the needs and interests of employers and subsequent providers of education and training to identify and select individuals. Within schools, educational assessments are used to determine the route a student takes through the differentiated curricula that are on offer, as well as to report on a student’s educational achievement either to the student herself, or to her parents or guardians. However educational assessments can also support learning, both by identifying obstacles to future learning, and by providing appropriate feedback to learners on what they need to do to improve.

For the purposes of this paper, I will use the term ‘evaluative’ to describe assessments that are designed to evaluate institutions and curricula, and which serve the purposes of accountability, and ‘summative’ to describe assessments that are used to certify student achievement or potential. I shall use the term ‘diagnostic’ for those assessments that provide information about the difficulties that a student is experiencing, and ‘formative’ for those that provide feedback to learners about how to go about improving.

The terms 'formative', 'diagnostic', 'summative' and 'evaluative' are generally used as if they describe kinds of assessments, but of course the outcomes of the same assessment might be used to serve more than one function. These terms are therefore not descriptions of kinds of assessment but rather of *the use to which information arising from the assessments is put* (Wiliam & Black, 1996).

Of course other authors have used these terms in different ways. In 1967 in an AERA monograph on evaluation, Michael Scriven distinguished between formative and summative evaluations (Scriven, 1967) but it was Bloom, Hastings & Madaus (1971) who were the first to extend the usage to its generally accepted current meaning. They defined as *summative evaluation tests* those assessments given at the end of units, mid-term and at the end of a course, which are designed to judge the extent of students' learning of the material in a course, for the purpose of grading, certification, evaluation of progress or even for researching the effectiveness of a curriculum. They contrasted these with "another type of evaluation which all who are involved—student, teacher, curriculum maker—would welcome because they find it so useful in helping them improve what they wish to do" (p117), which they termed 'formative evaluation'.

While this dichotomy seems perfectly unexceptionable, it appears to have had one serious consequence. There can be little doubt that significant tensions are created when the same assessments are required to serve multiple functions, and few authors appear to believe that a single system can function adequately to serve all four functions. At least two different systems are therefore required. It is my belief that the use of the terms 'formative' and 'summative' to describe a dichotomy between formative and diagnostic functions on the one hand, and summative and evaluative on the other, has influenced the decision about how these functions should be divided between the two assessment systems. In other words the 'formative'-'summative' distinction has produced a belief that one system should cater for the formative and diagnostic functions, and another should cater for the summative and evaluative functions. For the remainder of this paper, I will use quotation marks to denote traditional uses of the terms 'formative' and 'summative', so that 'formative' assessment encompasses diagnostic and formative assessment, while 'summative' assessment encompasses both evaluative and summative assessment.

Experience in many countries indicates that very few teachers are able or willing to operate parallel assessment systems—one designed to serve a 'summative' function and one designed to serve a 'formative' function. On this assumption, the incompatibility of 'formative' and 'summative' functions of assessment leads inevitably to one of two policy prescriptions. The first acknowledges the centrality of formative assessment to the task of teaching:

We shouldn't want it [a shift to formative assessment] because research shows how it improves learning (we don't need to be told that—it has to be true). We should want it because schools are places where learners should be learning more often than they are being selected, screened or tested in order to check up on their teachers. The latter are important; the former are why schools exist. (Silcock, 1998)

But the consequence of this would be to exclude teachers from the assessment of their students and to rely on external agencies for the summative (and evaluative) assessment of student learning. As well as having undesirable backwash effects, there is a large amount of empirical evidence that the political constraints that limit the time that can be made available for summative assessment would result in assessments that focus on only a small part of what is considered important.

The second alternative is to acknowledge the role that teachers have in summative and evaluative assessment, but then not to expect them to make any systematic use of the information they collect to improve their students' learning. In view of the huge gains in achievement that formative assessment appears to be able to bring about, this would be quite disastrous, both for education systems in general, and for learners in particular.

Therefore if we are serious about raising standards of performance in our schools and workplaces, then there is literally no alternative. Whatever a logical analysis of the problem suggests, rather than adopting entrenched positions on one side or other of the

debate, we must refuse to accept the incompatibility of ‘summative’ and ‘formative’ assessment. Instead, we must find ways of mitigating that tension, by whatever means we can. This paper is an attempt to contribute towards this end.

The assessment cycle

All four functions of assessment require that evidence of performance or attainment is elicited, is then interpreted, and as a result of that interpretation, some action is taken. Such action will then (directly or indirectly) generate further evidence leading to subsequent interpretation and action, and so on.

In order to investigate ways in which the tension between different functions of assessment can be mitigated, these three key phases—elicitation, inference and action—are investigated in turn below. Although there is no natural beginning or ending to the process of assessment, it is convenient to start with the elicitation of evidence.

Eliciting evidence

Before any inferences can be made, or actions taken, some evidence about the level of performance must be generated and observed. We can immediately distinguish between *purposive* and *incidental* evidence. Purposive evidence is that which is elicited as a result of a deliberate act by someone (usually the teacher) that is designed to provide evidence about a student’s knowledge or capabilities in a particular area. This most commonly takes the form of direct questioning (whether orally or in writing). Of course, this will not guarantee that if the student has any knowledge or understanding in the area being assessed, then evidence of that attainment will be elicited. One way of asking a question might produce no answer from the student, while a slightly different approach may elicit evidence of achievement. We can never be absolutely sure that we have exhausted all the possibilities, so that we can never be sure that the student does *not* know something, but some assessments will be better than others in this respect. Elsewhere, I have termed the extent to which an assessment can be relied upon to elicit evidence of the achievement of interest the *disclosure of the assessment* (Wiliam, 1992).

Incidental evidence, on the other hand, is evidence of achievement that is generated in the course of a teacher’s day-to-day activities, when the teacher notices that a student has some knowledge or capability of which she was not previously aware. Of course, the distinction between purposive and incidental evidence of achievement is not sharp. Rather we have a continuum in which either the purposive or the incidental end of the continuum is dominant. Direct questioning on a specific topic will be largely purposive, although the sensitive teacher will be alert to evidence about other topics that emerge from the student’s responses. An exploratory or investigative activity, on the other hand, because of its unpredictable course, will often produce largely incidental evidence, but of course the choice of the activity will have been made with a view to eliciting evidence of interest, and, to an extent is also purposive.

The distinction between purposive and incidental assessment is consistent with the idea that a certain amount of knowledge or capability exists within the individual being assessed, and evidence of which is generated more or less reliably by the assessment, which appears to be a widely held view (see for example Wiley & Haertel, 1996). However, the distinction between purposive and incidental assessment is also meaningful in a view of knowledge as being constructed during the assessment episode—for example if our assessment is an assessment *in*, rather than *of*, the zone of proximal development (Allal & Pelgrims Ducrey, 2000).

As well as the means by which it is generated, there are also differences in the *form* in which evidence is generated. Frequently, because of the concern to establish consistency across raters, only evidence that exists in some permanent form (as writing, artefacts, or on audio- or video-tape) has been relied upon in formal assessment settings, while *ephemeral evidence* has been largely ignored. However, as far as formative and diagnostic assessment is concerned, inter-

rater consistency is of secondary importance (see below), and ephemeral evidence can be an entirely appropriate form of evidence.

In terms of the tension between different functions of assessment the elicitation of evidence is probably the most problematic aspect of the assessment cycle. In any situation in which the primary purpose for the collection of data is summative, and particularly one in which the data is likely to be used evaluatively, there will always be a difficulty in eliciting data that can serve a formative or diagnostic function.

For example, part of the rhetoric of the current school inspection regime in England is that schools can 'improve through inspection'. Presumably, this occurs when the inspectors identify aspects of a school's practice that can be improved. However, given that it is widely perceived that the primary purpose of the inspections is not to improve schools, but merely to identify less successful ones, there is an inevitable tendency for the institution to ensure that during the period of inspection, all areas of potential difficulty are hidden from the sight of the inspectors (there have been reports of schools that hire large numbers of computers just for the week of an inspection!).

Similarly, in initial teacher education in the United Kingdom, it is common for the student's personal tutor to be involved in both the *development* and the *assessment* of competence. When the tutor visits the placement school to observe the student's professional practice, the tutor sees this as an opportunity for the student to discuss any difficulties with the tutor, so that support can be given. However, the student may well not wish to raise any difficulties with the tutor, just in case these are issues of which the tutor was not already aware, and might thus be taken into account in any summative assessment of performance.

In both these examples, a limited range of outcomes is taken to be a representative sample of all possible relevant outcomes. In a school inspection, each teacher will be inspected teaching particular lessons to particular classes, and the assumption that is made is that this is representative of that teacher's performance on any of the topics that they teach, with any class. Similarly, a student on a programme of initial teacher education will be observed by her or his tutor teaching a small number of lessons, and again, this is assumed to be representative of that teacher's performance with the same class at other times, on other topics, with other classes, and even in other schools.

In these examples, the tension between the evaluative/summative and diagnostic/formative functions of assessment in the elicitation of evidence arises principally because the information base on which the assessment is based is, or has the potential to be, a non-representative sample of the entire domain. If the inferences made on the basis of the sample of outcomes that were actually observed were the same as the inferences that would be made from any other sample from the same domain, then the tension would not arise. However the belief (on the part of the institution or the individual) that it is possible for the inferences based on the sample to be consistently more favourable than might otherwise be the case, because of the way that the sample is drawn, leads to an attempt to restrict the sample to those outcomes that support the more favourable outcomes.

For example, if we give a class of students a list of twenty words that we expect each student to be able to spell, and tell them that we will be testing them on the spelling of those twenty words on the following day, what can we conclude from the test results? The answer is, of course, not very much. If a student spelled all twenty words correctly, we know that, on at least one occasion, that she or he could spell those twenty words, but we are not justified in extending our inferences to other words, or even to other occasions. In this situation, a test tests only what a test tests. We have no justification for any inferences beyond the items actually assessed. This is of course, the fundamental problem of assessment—drawing inferences beyond the items actually assessed.

Typically, the approach taken is to regard the items included in a test as being a sample from a wider domain, and we use the proportion of the items in the test that are correctly answered as an estimate of the proportion of all possible items in the domain that the candidate would be able to do. This inference, of course, requires that the sample of items in our test is a random sample of all possible items from the domain. In other words, each item in the domain must

have the same chance of being selected, although in practice this is not the case, as Jane Loevinger pointed out 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 reason that the sampling approach fails in our spelling test is that far from being a random sample from a larger domain, the predictability of the selection from the domain means that the items selected come to constitute the whole of the domain of interest. The consequence of this is to send value messages to the students that only the twenty tested words matter (at least for the moment), with the social consequence that only those twenty words get learned. This provides a very small-scale example of some of the impacts of educational assessments in social settings, which can be analysed with Messick's two-facet model for validity argument.

Messick (1980) argues that validity argument can be regarded as the result of crossing two facets of assessment: the *basis* of the assessment (evidential versus consequential) with the *function* of the assessment (interpretation versus use), as shown in figure 1.

	result interpretation	result use
evidential basis	construct validity A	construct validity and relevance/utility B
consequential basis	value implications C	social consequences D

Figure 1: Messick's framework for the validation of assessments

The upper row of Messick's table relates to traditional conceptions of validity, while the lower row relates to the *consequences* of assessment interpretation and use. A sample of items is chosen originally for its ability to represent the domain of interest, to support generalisations within the domain of interest (cell A), or beyond it (cell B). However the consequence of making a particular selection is to convey messages about the values of particular aspects of the domain. Where this selection is some sense predictable, the danger is that those items actually selected come to be seen as more important than those that are not (cell C). The social consequences of this selection may (as in the case of our spelling test) include a change in the actions of those being assessed, to focus only on those items or aspects of the domain selected for assessment (cell D). This then weakens the relationship between the sample and the domain, so that in extreme case (such as our spelling test), the sample completely loses its ability to stand as a proxy for the original domain.

This analysis suggests three potential routes to alleviating or mitigating the tension between the formative and summative functions of assessment in the elicitation of evidence. The first, and one that is well-known in the literature, is to broaden the evidence base so that a greater range of outcomes are observed. Observing a larger sample of performances increases the reliability of the assessment, but much more importantly, reduces the possibility of systematic construct under-representation, so that the potential for biased inferences is reduced.

The second approach is to prevent the distortion of the sample of outcomes observed by denying the subject of the assessment the knowledge of when they are being assessed. This is the rationale behind the unannounced audit of a commercial organisation, or the snap inspection of a catering outlet, and—more sinister—the rationale behind Jeremy Bentham's *Panopticon* described by Foucault (1977). Whatever the ethics of using such a technique in prisons, or for food preparation, such an approach is unlikely to be ethically defensible in educational settings.

The third approach involves a shift of attention from quality control to quality assurance. Instead of assessing the quality of some aspect of performance, with the distortions that this can produce, the focus of assessment is lifted from the actual performance to the quality of systems of self-evaluation.

This is the approach used in the current cycle of Teaching Quality Assessment being carried out within Higher Education Institutions in England. Although the process is called 'Teaching Quality Assessment' or 'Subject Review', what is evaluated in this process is the quality of institutional processes of self-review. It is up to the institution to determine what are its aims in terms of teaching and learning, and the task of the assessors is limited to evaluating the extent to which the institution achieves those aims. Although assessors may observe teaching, their remit is to assess the congruence of the teaching observed with the stated aims. As one commentator wryly observed, "It doesn't matter if your teaching is useless as long as you can prove it"!

In the case of the school inspections described above, this would be accomplished by changing the focus of inspection away from the actual results achieved by students and the quality of teaching and towards the institution's own procedures for self-evaluation and review. The inspection of schools would focus on the capability of the school to keep its own practices under review. The quality of teaching would be observed, but only as far as was necessary to establish the extent to which the teachers and the managers within the school were aware of the quality of the students' learning, as advocated by, for example, MacBeath, Boyd, Rand and Bell (1996).

In this context, it is interesting to note that it is felt by the British Government that the demands of accountability for the expenditure of public money in universities can be met through a system of quality assurance, while at school level, nothing less than a system of quality control will do. This is in marked contrast to other systems, such as, for example, that of the autonomous region of Catalunya in Spain, where the regional government believes that the concerns of public accountability for its schools can be met through the establishment of a system of institutional self-review (Generalitat de Catalunya Departament d'Ensenyament, 1998).

In the case of the student-teacher on a programme of initial teacher education described above, such an approach could be implemented by changing the role of the college tutor from one of assessing teaching competence to assessing the quality of the student's own self-evaluation. This might involve observing the student teaching, but might well function more effectively if the role of the tutor was limited to discussing with the student the results of the student's own self-evaluation. In such a context, failure to find anything negative in one's own performance is not a neutral or positive outcome, but rather a highly negative one. To counter the possible tendency to invent minor or trivial difficulties, in order to be able to solve them easily, the tutor could question the student on the means by which the student became aware of aspects of her or his practice that could be improved. If it is necessary to distance the evaluation even further from the assessment of performance, the student could be required to keep a log of self-evaluations of performance, with the role of the tutor restricted to establishing that the log exists and is up-to-date, rather than being concerned directly with its assessment.

These are just two examples of how the principle of changing the focus of assessment, from the direct assessment of outcomes to a focus on the capability for self-evaluation, together with a broadening of the basis for assessment, could support the elicitation of evidence that would support both diagnostic/formative and evaluative/summative functions of assessment. It also emphasises that self-assessment, far from being an adjunct of effective formative assessment is actually at its core.

Interpreting evidence

Of course, the availability of evidence of attainment means nothing until it is interpreted. For most of the history of educational assessment, the predominant way of interpreting the result of assessments has been to compare the performance of an individual with that of a

more or less well-defined group of individuals. All such a comparison requires is that we are able to put the performance of the individuals into some kind of rank order, and it is very easy to place individuals in a rank order without having any clear idea of what they are in rank order of. Such norm- and cohort-referenced assessments are frequently based on ill-defined domains, which means they are very difficult to relate to future learning needs, and therefore cannot easily serve a diagnostic, let alone a formative function. However, even where the domain is well-defined, then norm- and cohort-referenced interpretations do not generally function formatively because they focus on how *well* someone has done, rather than on *what* they have done. A norm- or cohort-referenced interpretation of a test result would indicate how much better an individual needs to do, pointing to the *existence* of a ‘gap’ (Sadler, 1989), rather than giving any indication of how that improvement is to be brought about. In other words, telling an individual *that* they need to do better, rather than telling her or him *how* to improve (rather like telling an unsuccessful comedian to be funnier).

This would suggest that the interests of formative assessment would be adequately served by criterion-referenced assessments, but it is important to note that a criterion-referenced interpretation of the result of an assessment is a necessary, but not a sufficient condition for the assessment to function formatively. Knowing that a learner has difficulty with a particular aspect of the domain is more help to the teacher and the learner than just knowing that they need to do better on the domain, but for the assessment to function formatively, the interpretation of the assessment must be not just criterion-referenced, but interpreted in terms of learning needs. In other words, it must be not just diagnostic, but also (to use a much mis-used word) remedial. The essential condition for an assessment to function formatively is that it must provide evidence that can be interpreted in a way that suggests what needs to be done next to close the gap. The interpretation of the assessment outcome must include a recipe for future action and must be related to a developmental model of growth in the domain being addressed—in short, it must be based on a model of progression.

This has become very clear recently in our work at King’s with teachers on a research project designed to develop classroom assessment skills. Like most teachers in England and Wales, the teachers in the project are highly skilled at grading students’ work in terms of the grades used for reporting the results of the national school-leaving examination—perhaps best summed up by an external assessor who, in commenting on a student’s portfolio said, “This screams ‘D’ at me”! This is consistent with the findings from an international OECD study which found that the summative function of assessment predominates in the practice of many teachers (Black & Atkin, 1996).

However, teachers are much less skilled at saying how the work could be improved—in other words, what would need to be changed for this (say) ‘D’ piece of work to be worth a ‘C’? An assessment that is not based on a theory of learning might (conceivably) be able to function diagnostically but it cannot function formatively.

The pervasiveness of the summative role of assessment is illustrated by the following item taken from the algebra test developed for the Concepts in Secondary Mathematics and Science (CSMS) project (Hart, Brown, Kerslake, Küchemann, & Ruddock, 1985):

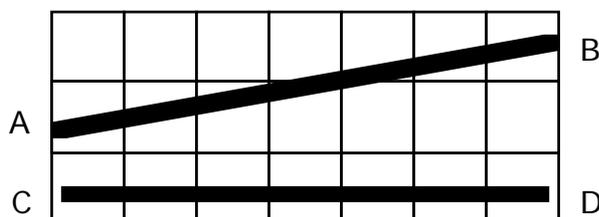
Simplify, if possible, $5a + 2b$

Many teachers regard this as an unfair item, since students are ‘tricked’ into simplifying the expression, because of the prevailing ‘didactic contract’ (Brousseau, 1984) under which the students assume that there is ‘academic work’ (Doyle, 1983) to be done. In other words ‘doing nothing’ cannot possibly be the correct answer because one does not get marks in a test without doing some work (much in the same way that ‘none of the above’ is never the correct option in a multiple choice test!). The fact that they are tempted to ‘simplify’ this expression in the context of a test question while they would not do so in other contexts means that this item may not be a very good question to use in a test serving a summative or evaluative purpose. However, such considerations do not disqualify the use of such an item for *diagnostic* purposes, because the fact that a student can be ‘tricked’ into simplifying this expression is relevant information for the teacher, indicating that the understanding of the basic principles of algebra is not secure.

Similar issues are raised by (so-called) ‘trick’ questions like:

1. Which of the following statements is true:

- (1) AB is longer than CD
- (2) AB is shorter than CD
- (3) AB and CD are the same length



2. Which of these two fractions is the larger?

$$\frac{3}{7} \qquad \frac{3}{11}$$

[Data from the CSMS study (Hart, 1981) indicates that while around 75% of 14-year-olds could choose which of two ‘ordinary’ fractions was the larger, only 14% correctly identified the larger from the two fractions above.]

Items such as these tend to exhibit highly undesirable psychometric properties, and are likely to be very poor indicators of the extent of an individual’s mastery of a wider domain. However, items such as these can provide profound insights into learning, and it is evidence of the pervasiveness of the evaluative/summative role of assessments that teachers regard such items as ‘unfair’ even in the context of classroom questioning.

Another illustration of the importance of separating the elicitation of evidence from its interpretation is provided by the following example from the development of the National Curriculum in England and Wales during the early 1990s.

In the first version of the national curriculum, the ‘attainment targets’ for mathematics and science were presented in terms of 296 statements of attainment, each of which was allocated to one of the ten levels of the national curriculum.

Many teachers devised elaborate record sheets that would allow them to indicate, for each statement of attainment, and for each student, whether it had been achieved. Originally, such a record sheet formed a formative function: it gave detailed criterion-referenced information on a student’s current attainment, and, just as importantly, what had not yet been attained. While some teachers did question the notion of progression inherent in the allocation of the statements of attainment to levels, most seemed happy to accept that the students’ next objectives were defined in terms of those statements just beyond the ‘leading edge’ of attained statements.

When a student produced evidence that indicated that she or he had partially achieved a statement (perhaps by demonstrating a skill in only a limited variety of contexts), then teachers would often not ‘tick off’ the statement, so that they would be reminded to re-evaluate the student’s performance in this area at some later date, perhaps in a different context. Since there are typically many opportunities to ‘re-visit’ a student’s understanding of a particular area, this seems a good strategy, given that a false-negative attribution (assuming that a student doesn’t know something they do, in fact, know) is, in an educational setting, likely to be far less damaging than a false-positive (assuming that they do know something they don’t).

However, schools were subsequently advised to derive the summative levels required in national curriculum assessment by the inflexible application of a formula. This immediately created a tension between diagnostic/formative and evaluative/summative functions. Where

teachers had left statements ‘unticked’ in order to prompt them to return to those aspects at a later date, students who had relatively complete understandings were often regarded as not having met the criterion, challenging the validity of the outcome as an accurate summation of the student’s achievement (and also, of course reducing the school’s ‘performance score’). In response to this, in the following year, teachers then stopped using the record sheets for diagnostic and formative functions, and started using them to record when the student had achieved a sufficient proportion of the domain addressed by the statement. The record sheets then were useful only for evaluative/summative purposes .

Here, the tension between diagnostic/formative and evaluative /summative functions arises because of the inflexible application of a mechanical rule for aggregation that has the effect of conflating the elicitation of evidence with its interpretation. The distorting effect of the summative assessment could have been mitigated if, instead of using an algorithmic formula, the summative assessment had involved a process of *re-assessment* of the original evidence.

More generally, the tension between formative and summative functions of assessment *may* therefore be ameliorated by separating the elicitation of evidence from its interpretation, and to interpret evidence differently for different purposes.

The question that then arises is which function should serve as the foundation:

It is possible to build up a comprehensive picture of the overall achievements of a pupil by aggregating, in a structured way, the separate results of a set of assessments designed to serve a formative purpose. However, if assessments were designed only for summative purposes, then formative information could not be obtained, since the summative assessments occur at the end of a phase of learning and make no attempt at throwing light on the educational history of the pupil. It is realistic to envisage, for the purpose of evaluation, ways of aggregating the information on individual pupils into accounts of the success of a school, or LEA [Local Educational Authority] in facilitating the learning of those for whom they are responsible; again the reverse is an impossibility (National Curriculum Task Group on Assessment and Testing, 1988 ¶25).

Since it is impossible to disaggregate summary data, and relatively easy to aggregate fine-scale data, this suggests that some mitigation of the tension between formative and summative assessment may be achieved by making the formative/diagnostic function paramount in the elicitation of evidence, and by interpreting the evidence in terms of learning needs in the day-to-day work of teaching. When it is required to derive a summative assessment, then rather than working from the already interpreted information, the teacher goes back to the original evidence, ignoring those aspects (such as ‘trick questions’) that are relevant for the identification of learning needs, but less relevant for determining the overall level of achievement.

Finally the use of the same assessments to serve both summative and evaluative functions can blind us to the fact that these functions have very different requirements. The use of assessments for evaluative purposes routinely requires that the results of individuals are aggregated—usually down to a single number or grade for each individual. However, the multi-dimensionality of performance—even in the most narrowly-defined domains—means that effective summative assessment also needs to be multidimensional. Therefore, if we must have two systems, this suggests that it would be better to separate the evaluative function from the others. In other words, we would have one (external) system serving the evaluative function, and another system, driven mainly by teachers’ own assessments of their students, serving the summative, diagnostic and formative functions. How this might be achieved is discussed below.

Action

In terms of traditional models of validity, the assessment process reaches an end when assessment outcomes are interpreted. While there may be some actions contingent on the outcomes, they tend to follow directly and automatically, as a result of previous validation studies. Students who achieve a given score on the SAT are admitted into college because the score is taken to indicate that they have the necessary aptitude for further study. In other words, the primary focus of validity is on the meanings and significance (Bechtoldt, 1959) of

the assessment outcomes. One essential requirement here is that the meanings and significance of the assessment outcomes must be widely shared. The same score must be interpreted in similar ways for different individuals. The value implications and the social consequences of the assessment, while generally considered important, are often not considered as aspects of validity at all (Madaus, 1988) and even where they are, are generally subsidiary to the consistency of interpretations. In a very real sense, therefore, summative and evaluative assessments are validated primarily with respect to their *meanings*.

For diagnostic and formative assessments, however, it is the learning that is caused as a result of the assessment that is paramount. If different teachers elicit different evidence from the same individual, or interpret the same evidence differently, or even if they make similar interpretations of the data, but then take different actions, then this is relatively unimportant, in that what really matters is whether the result of the assessment is successful learning. In this sense, formative assessments are validated primarily with respect to their *consequences*.

Discussion

To sum up, in order to serve a formative function, an assessment must yield evidence that, interpreted in terms of a theory of learning, indicates the existence of a gap between actual and desired levels of performance (the diagnostic function), and suggests actions that are in fact successful in closing the gap (the formative function). Crucially, an assessment that is *intended* to be formative (ie has a formative *purpose*) but does not, ultimately have the intended effect (ie lacks a formative *function*), is not formative.

Inevitably, these requirements are often at variance with the needs of assessments whose primary purpose is to attest to the level of knowledge, skill, aptitude, capability or whatever that an individual possesses. Formative, diagnostic, summative and evaluative assessments *do* serve conflicting interests, but this is not to say that they are incompatible. The consequences of accepting that the same assessments cannot serve both formative and summative functions results either in a pedagogy in which teachers make no systematic attempt to find out what her students are learning, or to a situation in which all important decisions about the achievement of students are made without reference to the person who probably knows most about that individual (as one commentator observed, “Why rely on an out-of-focus snapshot taken by a total stranger?”).

In this paper, I have suggested that as separation of the elicitation of evidence from its interpretation, and the consequent action may help in this undertaking. Where the same assessment is to serve both formative and summative purposes, the basis of the assessment must be broad, and must, as far as possible, not be predictable (at least not to the extent that those being assessed can ignore certain parts of the domain because they know that these will not be assessed). Consideration should also be given to changing the focus of the assessment from a quality control orientation, where the emphasis is on the external assessment as the measurement of quality, to a quality assurance orientation, where the emphasis is on the evaluation of internal systems of self-assessment, self-appraisal or self-review.

Once evidence is elicited, it should be interpreted differently for different purposes. For evaluative and summative purposes the primary focus should be on synopsis. The available evidence should be interpreted in order to provide the best summary of the individual’s achievement or potential in a particular domain, essentially through some sort of random (or stratified-random) sample of the domain. The results on ‘trick questions’ are given relatively little weight, or are ignored completely.

For diagnostic and formative purposes, however, the focus should be on learning. Some items are much more important than others, since they have a greater propensity to disclose evidence of learning needs. In particular, the results on some sorts of ‘trick questions’ can be especially significant, because they can point clearly to learning needs that were not previously clear and may even suggest the appropriate action to take. In this context, it is important to note that once the data has been interpreted for one purpose, it cannot easily serve another. The aggregation of fine-scaled data from diagnostic or formative assessments

is best achieved not by a process of aggregation of already interpreted data, but rather by a re-assessment, for a different purpose, of the original evidence.

Summative assessments are best thought of as retrospective. The vast majority of summative assessments in education are assessments of what the individual has learnt, knows, understands or can do. Even where the assessments are used to predict future performance, this is done on the basis of *present* capabilities, and assessments are validated by the consistency of their meanings. In contrast formative assessments can be thought of as being *prospective*. They must contain within themselves a recipe for future action, whose validity rests in their capability to cause learning to take place, which is, after all, the main purpose of education.

Finally, this analysis suggests that the evaluative function of assessment is best undertaken by a separate system from that designed to contribute to summative, diagnostic and formative functions. Where the same system has to serve both evaluative and summative functions there is always the danger of the narrowing of the curriculum caused by 'teaching to the test'. Even school-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 can be provided by using a large number of tasks that cover the entire curriculum, with each student randomly assigned to take a small number of these tasks. The task would not provide an accurate measure of that student's achievement (because of the degree of student-task interaction) but the mean score for all students would be a highly 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 overall levels achieved on the evaluative assessments could then be used to define an 'envelope' of overall scores to which the school would have to adjust its internal grades, thus assuring the comparability of the summative assessments across schools.

References

- Allal, L. & Pelgrims Ducrey, G. (2000). Assessment of—or in—the zone of proximal development. *Learning and Instruction*, **10**(2), 137-152.
- Bechtoldt, H. P. (1959). Construct validity: a critique. *American Psychologist*, **14**, 619-629.
- Black, P. J. & Atkin, J. M. (Eds.). (1996). *Changing the subject: innovations in science, mathematics and technology education*. London, UK: Routledge.
- Black, P. J. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles Policy and Practice*, **5**(1), 7-73.
- Bloom, B. S.; Hastings, J. T. & Madaus, G. F. (Eds.). (1971). *Handbook on the formative and summative evaluation of student learning*. New York, NY: McGraw-Hill.
- Boaler, J.; Wiliam, D. & Brown, M. L. (2000). Students' experiences of ability grouping—disaffection, polarisation and the construction of failure. *British Educational Research Journal*, **27**(4).
- Brousseau, G. (1984). The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics. In H.-G. Steiner (Ed.) *Theory of mathematics education: ICME 5 topic area and miniconference* (pp. 110-119). Bielefeld, Germany: Institut für Didaktik der Mathematik der Universität Bielefeld.
- Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, **58**(4), 438-481.
- Daugherty, R. (1995). *National curriculum assessment: a review of policy 1987-1994*. London, UK: Falmer Press.
- Doyle, W. (1983). Academic work. *Review of Educational Research*, **53**(2), 159-199.
- Foucault, M. (1977). *Discipline and punish* (Sheridan-Smith, A M, Trans.). Harmondsworth, UK: Penguin.
- Generalitat de Catalunya Departament d'Ensenyament (1998). *Avaluació interna de centres*. Barcelona, Spain: Generalitat de Catalunya Departament d'Ensenyament.
- Gewirtz, S.; Ball, S. J. & Bowe, R. (1995). *Markets, choice and equity in education*. Buckingham, UK: Open University Press.

- Goldstein, H. (1996). Introduction. *Assessment in Education: Principles Policy and Practice*, 3(2), 125-128.
- Hart, K. M. (Ed.) (1981). *Children's understanding of mathematics: 11-16*. London, UK: John Murray.
- Hart, K. M.; Brown, M. L.; Kerslake, D.; Küchemann, D. & Ruddock, G. (1985). *Chelsea diagnostic mathematics tests*. Windsor, UK: NFER-Nelson.
- Kellner, P. (1997). Hit-and-miss affair. *Times Education Supplement*, 23.
- Linn, R. L. (1994) *Assessment-based reform: challenges to educational measurement*. Paper presented at Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service.
- Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, 72(2), 143-155.
- MacBeath, J.; Boyd, B.; Rand, J. & Bell, S. (1996). *Schools speak for themselves: towards a framework for self-evaluation*. London, UK: National Union of Teachers.
- Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.) *Critical issues in curriculum: the 87th yearbook of the National Society for the Study of Education (part 1)* (pp. 83-121). Chicago, IL: University of Chicago Press.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35(11), 1012-1027.
- Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist*, 22(2), 155-175.
- Reay, D. & Wiliam, D. (1999). I'll be a nothing: structure, agency and the construction of identity through assessment. *British Educational Research Journal*, 25(3), 343-354.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 145-165.
- Scriven, M. (1967). *The methodology of evaluation*. Washington, DC: American Educational Research Association.
- Wiley, D. E. & Haertel, E. H. (1996). Extended assessment tasks: purposes, definitions, scoring and accuracy. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: promises, problems and challenges* (pp. 61-89). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wiliam, D. (1992). Some technical issues in assessment: a user's guide. *British Journal for Curriculum and Assessment*, 2(3), 11-20.
- Wiliam, D. (1995a). The development of national curriculum assessment in England and Wales. In T. Oakland & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 157-185). Boston, MA: Kluwer Academic Publishers.
- Wiliam, D. (1995b). Combination, aggregation and reconciliation: evidential and consequential bases. *Assessment in Education: Principles Policy and Practice*, 2(1), 53-73.
- Wiliam, D. & Black, P. J. (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, 22(5), 537-548.