Integrating summative and formative functions of assessment[1]

Dylan Wiliam
King's College London

0. Abstract

This paper suggests ways in which the tension between the summative and formative functions of assessment might be ameliorated. Following Messick, it is suggested that the consideration of social consequences is essential in the validation of assessments, and it is argued that most assessments are interpreted not with respect to norms and criteria, but by reference to constructs shared amongst communities of assessors. Formative assessment is defined as all those activities undertaken by teachers and learners which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged, and is characterised by four elements: questioning, feedback, sharing quality criteria and student self-assessment. Assessment is then considered as a cycle of three phases (eliciting evidence, interpreting evidence, taking action), and ways in which the tensions between summative and formative functions of assessment can be ameliorated are considered for each of these phases.

1. Introduction

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 find ways of selecting 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, arguably the most important function that educational assessments serves is in supporting learning:

> 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. (Peter Silcock, 1998, personal communication)

Traditionally, the informal day-to-day use of assessment within classrooms to guide learning has received far less attention than the more formal uses, and to the extent that it has been discussed at all, it has tended to be discussed as an aspect of pedagogy or instructional design. However, within the past ten years, there has been a recognition of the need to integrate (or at least align) the routines of informal classroom assessment with more formal assessment practices. It has become conventional to describe these two kinds of assessment as formative and summative assessment respectively, but it is important to note in this context that the terms 'formative' and 'summative' do not describe assessments—the same assessment might be used both formatively and summatively—but rather are descriptions of *the use to which information arising from the assessments is put*.

In this paper, my aim is to attempt to show how these two functions of assessment might be integrated, or at least aligned. In section 2 I outline some theoretical foundations related to the summative functions of assessment, and suggest that the role of professional judgement in all assessments is actually much greater than is usually supposed. In section 3, I sketch out the results of a recent substantial review of the research on the effectiveness of formative assessment, using a framework that integrates the role of the teacher and of the student. In section 4, I propose a series of conflicts between the summative and formative functions of assessment and suggest ways these tensions may be ameliorated or softened.

---

[1] Keynote address to the European Association for Educational Assessment; Prague: Czech Republic, November 2000

2. Summative Assessment

If a teacher asks a class of students to learn twenty number bonds, and later tests the class on these bonds, then we have a candidate for what Hanson (1993) calls a 'literal' test. The inferences that the teacher can justifiably draw from the results are limited to exactly those items that were actually tested. The students knew which twenty bonds they were going to be tested on, and so the teacher could not with any justification conclude that those who scored well on this test would score well on a test of different number bonds.

However, such kinds of assessment are rare. Generally, an assessment is "a representational technique" (Hanson, 1993 p19) rather than a literal one. Someone conducting an educational assessment is generally interested in the ability of the result of the assessment to stand as a proxy for some wider domain. This is, of course, an issue of validity—the extent to which particular inferences (and, according to some authors, actions) based on assessment results are warranted.

In the predominant view of educational assessment it is assumed that the individual to be assessed has a well-defined amount of knowledge, expertise or ability, and the purpose of the assessment task is to elicit evidence regarding the amount or level of knowledge, expertise or ability (Wiley & Haertel, 1996). This evidence must then be interpreted so that inferences about the underlying knowledge, expertise or ability can be made. The crucial relationship is therefore between the task outcome (typically the observed behaviour) and the inferences that are made on the basis of the task outcome. Validity is therefore not a property of tests, nor even of test outcomes, but a property of the inferences made on the basis of these outcomes. As Cronbach noted forty-five years ago, "One does not validate a test, but only a principle for making inferences" (Cronbach & Meehl, 1955 p297).

Traditional assessments, typically written examinations and standardised tests, can assess only a small part of the learning of which they are claimed to be a synopsis. In the past, this has been defended on the grounds that the test is a random sample from the domain of interest, and that therefore the techniques of statistical inference can be used to place confidence intervals on the estimates of the proportion of the domain that a candidate has achieved, and indeed, the correlation between standardised test scores and other, broader measures of achievement are often quite high.

However, after a moment's reflection, it is clear that the contents of standardised tests and examinations are not a random sample from the domain of interests. In particular, timed written assessments can assess only limited forms of competence, and teachers are quite able to predict which aspects of competence will be assessed. Especially in 'high-stakes' assessments, therefore, there is an incentive for teachers and students to concentrate on only those aspects of competence that are likely to be assessed. Put crudely, we start out with the intention of making the important measurable, and end up making the measurable important. The effect of this has been to weaken the correlation between standardised test scores and the wider domains for which they are claimed to be an adequate proxy. This provides a vivid demonstration of the truth of Goodhart's law.

2.1 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 British 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 attending the school might have been well below the national average 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 (the 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).

Concerns with the 'manipulability' of traditional tests (ie the ability to improve students' scores on the tests without increasing their performance on the domain of which the test purports to be a sample) has led to increasing interest in the use of 'authentic' or 'performance' assessment (Resnick & Resnick, 1992). If we want students to be able to apply their knowledge and skills in new situations, to be able to investigate relatively unstructured problems, and to evaluate their work, tasks that embody these attributes must form part of the formal assessment of learning—a test is valid to the extent that one is happy for teachers to teach towards the test (Wiliam, 1996a).

These problems of manipulability arise because educational assessments are used in essentially social situations. As with money supply and inflation, placing great emphasis on the correlates of educational achievement, such as tests, changes the relationship between the index and what it is taken to be an index of. Authors differ on whether these concerns with the social consequences of educational assessments should be regarded as part of validity. Some, such as Madaus (1998), have argued that the *impact* of an assessment is conceptually quite distinct from its validity. Others, notably Samuel Messick, have argued that consideration of the consequences of the use of assessment results is central to validity argument. In his view, "Test validation is a process of inquiry into the adequacy and appropriateness of interpretations and actions based on test scores" (Messick, 1989 p31).

Messick argues that this complex view of validity argument can be regarded as the result of crossing 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 validity enquiry

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. One of the consequences of the interpretations made of assessment outcomes is that those aspects of the domain that are assessed come to be seen as more important than those not assessed, resulting in implications for the values associated with the domain. For example, if open-ended and or practical work in a subject is not formally assessed, this is often interpreted as an implicit statement that such aspects of the subject are less important than those that are assessed. One of the social consequences of the use of such limited assessments is that teachers then place less emphasis on (or ignore completely) those aspects of the domain that are not assessed.

The incorporation of authentic assessment into 'high-stakes' assessments such as school-leaving and university entrance examinations can be justified in each of the facets of validity argument identified by Messick.

A Many authors have argued that assessments that do not include authentic tasks do not adequately represent the domain. This is an argument about the evidential basis of result interpretation (such an assessment would be said to under-represent the construct of the subject being assessed).

B It might also be argued that the omission of authentic tasks reduces the ability of assessments to predict a student's likely success in advanced studies in the subject, which would be an argument about the evidential basis of result use.

C It could certainly be argued that leaving out authentic tasks would send the message that such aspects of the subject are not important, thus distorting the values associated with the domain (consequential basis of result interpretation).

D Finally, it could be argued that unless authentic tasks were incorporated into the assessment, then teachers would not teach, or place less emphasis on, these aspects (consequential basis of result use).

However, if authentic tasks are to feature in formal 'high-stakes' assessments, then users of the results of these assessments will want to be assured that the results are sufficiently reliable. The work of Linn and others (see, for example, Linn & Baker, 1996) has shown that in the assessment of authentic tasks, there is a considerable degree of task variability. In other words, the performance of a student on a specific task is influenced to a considerable degree by the details of that task, and in order to get dependable results, we need to assess students' performance across a range of authentic tasks (Shavelson, Baxter, & Pine, 1992), and even in mathematics and science, this is likely to require at least six tasks. Since it is hard to envisage any worthwhile authentic tasks that could be completed in less than two hours, the amount of assessment time that is needed for the dependable assessment of authentic tasks is considerably greater than can reasonably be made available in formal external assessment. The only way, therefore, that we can avoid the narrowing of the curriculum that has resulted from the use of timed written examinations and tests is to conduct the vast majority of even high-stakes assessments in the classroom.

One objection to this is, of course, that such extended assessments take time away from learning. There are two responses to this argument. The first is that authentic tasks are not just assessment tasks, but also learning tasks; students learn in the course of undertaking such tasks and we are therefore assessing students' achievement not at the start of the assessment (as is the case with traditional tests) but at the end —the learning that takes place during the task is recognised. The other response is that the reliance on traditional assessments has so distorted the educational process leading up to the assessment that we are, in a very real sense, "spoiling the ship for a half-penny-worth of tar". The ten years of learning that students undertake in developed countries during the period of compulsory schooling is completely distorted by the assessments at the end. Taking (say) twelve hours to assess students' achievement in order not to distort the previous *thousand* hours of learning in (say) mathematics or the mother-tongue seems like a reasonable compromise.

Another objection that is often raised is the cost of marking such authentic tasks. The conventional wisdom in many countries is that, in high-stakes settings, the marking of the work must be conducted by more than one rater. However, the work of Linn cited earlier shows that rater variability is a much less significant source of unreliability than task variability. In other words, if we have a limited amount of time (or, what amounts to the same thing, money) for assessing students' work, results would be more reliable if we had six tasks marked by a single rater than three tasks each marked by two raters. The question that remains, then, is who should do the marking?

The answer to this question appears to depend as much on cultural factors as on any empirical evidence. In some countries (eg England, and increasingly over recent years, the United States) the distrust of teachers by politicians is so great that involving teachers in the formal assessment of their own students is unthinkable. Any yet, in many other countries (eg Norway, Sweden) teachers are responsible not just for

determination of their students' results in school leaving examinations, but also for university entrance. Given the range of ephemeral evidence that is likely to be generated by authentic tasks, and the limitations of even authentic tasks to capture all the learning achievements of students, the arguments for involving teachers in the summative assessment of their students seem compelling. As one commentator remarked: 'Why rely on an out-of-focus snapshot taken by a total stranger?'

However, if teachers are to be involved in the summative assessments of their students, it is necessary to be clear about how (and perhaps more importantly, with respect to what?) the validation should be conducted. That is an issue of the 'referent' of the assessment.

2.2 Norm-referenced and criterion-referenced assessment

For most of the history of educational assessment, the primary method of interpreting the results of assessment has been to compare the results of a specific individual with a well-defined group of other individuals (often called the 'norm' group), the best known of which is probably the group of college-bound students (primarily from the north-eastern United States) who in 1941 formed the original norm group for the Scholastic Aptitude Test.

Norm-referenced assessments have been subjected to a great deal of criticism over the past thirty years, although much of this criticism has generally overstated the amount of norm-referencing actually used in standard setting, and has frequently confused norm-referenced assessment with *cohort*-referenced assessment (Wiliam, 1996b).

However, the real problem with norm-referenced assessments is that, as Hill and Parry (1994) have noted in the context of reading tests, it is very easy to place candidates in rank order, without having any clear idea of what they are being put in rank order *of* and it was this desire for greater clarity about the relationship between the assessment and what it represented that led, in the early 1960s, to the development of 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 (Popham, 1980). 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, *op cit*).

However, any criterion, no matter how tightly drawn, has a degree of *plasticity*[2]. For example, the national curriculum for English in England and Wales requires that in order to achieve level 7 (the standard achieved by the average sixteen-year-old and by 10% of 14-year-olds) "Paragraphing and correct punctuation are used to make the sequence of events or ideas coherent and clear to the reader" (Department for Education & Welsh Office, 1995 p31). However, it is doubtful whether any teachers, let alone school students, could punctuate the following to make the meaning of the following sentence "clear to the reader"[3]:

paul where john had had had had had had had had had had had a clearer meaning

This problem is not confined to subjects such as language, where domain definition is notoriously difficult. Even in mathematics—arguably the subject in which domain definition is most straightforward—difficulties abound. For example, if we set a criterion such as "can compare two fraction to find which is the larger", the question that arises immediately is, "Which two fractions?"

Where the fractions have a common denominator but a different numerator, the facility is likely to be quite high (Hart, 1981 found facilities of the order of 95% for 14-year-olds), and even for fractions such as and the facility remains high (Hart found approximately 75%). However, when the fractions share a common *numerator* but a different *denominator*, then the facility is much lower (Hart found that only about 15% of 14-year-olds could do such items). Which fractions we would choose to exemplify this criterion would depend on what we want the criterion to do. Norms thus intrude into our criterion-referenced assessments in two ways. In the first place, they determine which criteria which we find interesting. For example, 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. In the second place, norms are used to determine what our criteria actually mean. As Angoff (1974)

[2] I am grateful to Jon Ogborn for suggesting this term.

[3] One way of punctuating this to being out the meaning is as follows. Paul, where John had had 'had', had had 'had had'. 'Had had' had had a clearer meaning.

pointed out "lurking behind the criterion-referenced evaluation, perhaps even responsible for it, is the norm-referenced evaluation" (p4).

Even if it were possible to define performance domains unambiguously, it is by no means clear that this would be desirable. Greater and greater specification of assessment objectives results in a system in which students and teachers are able to predict quite accurately what is to be assessed, and creates considerable incentives to narrow the curriculum down onto only those aspects of the curriculum to be assessed (Smith, 1991). The alternative to "criterion-referenced hyperspecification" (Popham, 1994) is to resort to much more general assessment descriptors which, because of their generality, are less likely to be interpreted in the same way by different assessors, thus re-creating many of the difficulties inherent in norm-referenced assessment. Thus neither criterion-referenced assessment nor norm-referenced assessment provides an adequate theoretical underpinning for authentic assessment of performance.

The ritual contrasting of norm-referenced and criterion-referenced assessments, together with more or less fruitless arguments about which is better, has tended to reinforce the notion that these are the only two kinds of inferences that can be drawn from assessment results. However the oppositionality between norms and criteria is only a theoretical model, which, admittedly, works well for certain kinds of assessments. But like any model, it has its limitations. My position is that the contrast between norm and criterion-referenced assessment represents the concerns of, and the kinds of assessments developed by, psychometricians and specialists in educational measurement. Beyond these narrow concerns there are a range of assessment events and assessment practices, characterised by authentic assessment of performance, that are routinely interpreted in ways that are not faithfully or usefully described by the contrast between norm and criterion-referenced assessment (this is particularly strong in the traditions of school examinations in European countries, and by the day-to-day practices of teachers).

Such authentic assessments have only recently received the kind of research attention that has for many years been devoted to standardised tests for selection and placement, and, indeed, much of the investigation that has been done into authentic assessment of performance has been based on a 'deficit' model, by establishing how far, say, the assessment of portfolios of students' work, falls short of the standards of reliability expected of standardised multiple-choice tests.

However, if we adopt a phenomenological approach, then however illegitimate these authentic assessments are believed to be, there is still a need to account for their widespread use. Why is it that the forms of assessment traditionally used in Europe have developed the way they have, and how is it that, despite concerns about their 'reliability', their usage persists?

What follows is a different perspective on the interpretation of assessment outcomes—one that has developed not from an a priori theoretical model but one that has emerged from observation of the practice of assessment within the European tradition.

2.3 Construct-referenced assessment

The model for the interpretation of assessment results that I wish to propose is illustrated by 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. 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 a marker for some fixed period of time.

The innovative feature of such assessment is that there is no need to prescribe learning outcomes. In that they are defined at all, they are 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 neither is it subjective, relying on the judgement or opinion of an individual. When there is agreement it arises from a degree of *intersubjectivity*. To put it crudely, it is not necessary for the raters (or anybody else) to know what they are doing, only that they do it right. Because the assessment system relies on the existence of a construct (of what it means to be competent in a particular domain) being shared among a community of assessors, I have proposed elsewhere that such assessments are best described as 'construct-referenced' (Wiliam, 1994). Another example of such a construct-referenced assessment is the educational assessment with perhaps the highest stakes of all—the PhD.

In most countries, the PhD is awarded as a result of an examination of a thesis, usually involving an oral examination. As an example, the University of London regulations provide what some people might regard as a 'criterion' for the award. In order to be successful the thesis must make "a contribution to original knowledge, either by the discovery of new facts or by the exercise of critical power". The problem is what is to count as a new fact? The number of words in this paper is, currently, I am sure, not known to anyone, so a simple count of the number of words in this paper would generate a new fact, but there is surely not a university in the world that would consider a simple word count worthy of a PhD.

The 'criterion' given creates the impression that the assessment is criterion-referenced one, but in fact, the criterion does not admit of an unambiguous meaning. To the extent that the examiners agree (and of course this is a moot point), they agree not because they derive similar meanings from the regulation, but because they already have in their minds a notion of the required standard. Of course, there are many instances where there is considerable evidence that no such collective agreement exists (see, for example, Morgan, 1998) but this in no way invalidates the argument I am advancing. Rather it points to situations where the community of practice has not yet been established, or, as is frequently the case, where there are two communities of practice, each relatively coherent, but differing in their values. The consistency of such assessments depend on what (Polanyi, 1958) called *connoisseurship*, but perhaps might be more useful regarded as the membership of a community of practice (Lave & Wenger, 1991).

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 learners in the same way as they are interpreted by teachers. For example, the national curriculum for English in England and Wales (*op cit*) 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 my view it is more appropriately thought of as 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. This is in some ways similar to the distinction between learning a first and a second language. When learning a second language, one is often (although of course, not always) learning a new label for an existing construct. The use of explicit criteria presupposes such a model of learning, in which students know what 'being sensitive' looks and feels like, and have only to reproduce the required behaviour. However, for most learners, it seems more plausible that the problem with failing to show sensitivity is that they have no clear idea of what the label is labelling. In a very real sense, therefore, the words cannot carry the meaning required. 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 over 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 or definitions 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).

This is not to say that such maxims are of no use—indeed quite the reverse. Maxims can be very useful in providing a focus for the negotiation of meanings, and once that has been done, the same maxims provide useful labels and a shorthand for describing quality. However, to reinforce Polanyi's point, the maxims do not themelves establish the definition of quality.

This notion of definition provides a touchstone for distinguishing between criterion- and construct-referenced assessment. 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.

2.4 How to do things with assessments

In the 1955 William James lectures J L Austin, discussed two different kinds of 'speech acts'—illocutionary and perlocutionary (Austin, 1962). Illocutionary speech acts are *performative*—by their

mere utterance they actually do what they say, creating social facts (Searle, 1995). For example by saying 'I invite you to a party' the social fact of the invitation is created by the utterance. The verdict of a jury in a trial is also an illocutionary speech act, 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. The other commonly quoted example of a 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") creates the social fact of the marriage.

Searle himself illustrates the idea of social facts by an interview between a baseball umpire and a journalist who was trying to establish whether the umpire believed his calls to be 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.

The umpire's calls bring into being social facts because the umpire is *authorised* (in the sense of having both the power, and that use of power being regarded as legitimate) to do so. The extent to which these judgements are seen as warranted ultimately resides in the degree of trust placed by those who use the results of the assessments (for whatever purpose) in the community of practice making the decision about membership (Wiliam, 1996b).

In my view a great deal of the confusion that currently surrounds educational assessments—particularly those in the European tradition—arises from the confusion of these two kinds of speech acts. Put simply, most educational assessments are treated as if they were perlocutionary speech acts, whereas in my view they are more properly regarded as illocutionary speech acts.

These difficulties are inevitable as long as the assessments are required to perform a perlocutionary function, making warrantable statements about the student's previous performance, current state, or future capabilities. Attempts to 'reverse engineer' assessment results in order to make claims about what the individual can do have always failed, because of the effects of compensation between different aspects of the domain being assessed.

However, many of the difficulties raised above diminish considerably if the assessments are regarded as serving an *illocutionary* function. To see how this works, it is instructive to consider the assessment of the PhD discussed above

Although technically, the award is made by an institution, the decision to award a PhD is made on the recommendation of examiners. In some countries, this can be the judgement of a single examiner, while in others it will be the majority recommendation of a panel of as many as six. The important point for our purposes is that the degree is the result of a speech act of an institution, acting on the recommendation (another speech act!) of examiners. The perlocutionary content of this speech act is negligible, because, if we are told that someone has a PhD, there are very few inferences that are warranted. In other words, when we ask "What is it that we know about what this person has/can/will be able to do once we know they have a PhD?" the answer is "Almost nothing" simply because PhD theses are so varied. Instead, the award of a PhD is better thought of not as an assessment of aptitude or achievement, or even as a predictor of future capabilities, but rather as an illocutionary speech act that *inaugurates an individual's entry into a community of practice*.

This goes a long way towards explaining the lack of concern about measurement error within the European tradition of examining. When a jury makes a decision the person is either guilty or not guilty, irrespective of whether they actually committed the crime—there is no 'measurement error' in the verdict. The speech act of the jury in announcing its verdict creates the social fact of someone's guilt until that social fact is revoked by a subsequent appeal, (creating a new social fact). In the European tradition of examining, examination authorities create social facts by declaring the results of the candidates, provided that the community of users of assessment results accept the authority of the examining body to create social facts. That is why, in a very real sense, that as far as educational assessment is concerned, there is no measurement error in Europe!

The foregoing theoretical analysis creates, I believe, a framework for the validation of summative assessment of student performance, whether conducted by an external examination body or by the teachers of those students themselves. Such assessments are construct-referenced assessments, validated by the extent to which the community of practice agrees that the student's work has reached a particular implicit standard. Achievement of this standard should not be interpreted in terms of a range of competences that the student had, has, or is likely to achieve at some point in the future, but instead as a

statement that the performance is adequate to inaugurate the student into a particular community of practice.

3. Feedback and Formative Assessment

Many authors have argued that formative assessment—that is in-class assessment of students by teachers in order to guide future learning (ie assessment *for* learning rather than assessment *of* learning)—is an essential feature of effective pedagogy. However, empirical evidence for its utility has, in the past, been widely distributed and rather difficult to locate. For example, two review articles (Crooks, 1988; Natriello, 1987) published at approximately the same time, and reviewing substantially the same field, included between them 323 references, of which only 9 were common to both. The approach adopted by Black and Wiliam, in their review of work in this area since the papers of Natriello and Crooks, was to review the contents of each issue of the 76 most relevant journals published from 1988 to 1997, yielding a total of 681 potentially relevant references, of which 251 were included in their final review (Black & Wiliam, 1998a).

Black and Wiliam found that research studies from all over the world, across a range of subjects, and conducted in primary, secondary and tertiary classrooms, with ordinary teachers, found consistent effect sizes of the order of 0.7. This is sufficient to raise the achievement of an average student to that of the upper quartile, or, expressed more dramatically, to raise the performance of an 'average' country (such as New Zealand, the United States, England or Germany) in the recent international comparisons of mathematics performance (TIMSS) to fifth place after Singapore, Japan, Taiwan and Korea.

There is also evidence, by its nature more tentative, that the current 'state of the art' in formative assessment is not well developed (Black & Atkin 1996, Black & Wiliam, 1998b), and therefore considerable improvements in learning can be achieved by effective implementation of formative assessment.

This section summarises some of the findings of that review, and elaborates some theoretical notions that are useful in developing an integrated assessment framework that encompasses both formative and summative functions (see section 4).

The term 'feedback' has its origins in engineering systems theory, and was originally used to describe a feature of a system that allows information regarding some aspect of the output of a system to be 'fed back' to the input in order to influence future outputs. Typically, this 'aspect of the output' is the level of some measurable attribute, and the process of feedback involves comparing the level of the output attribute either with some predetermined reference level, or (more commonly in engineering systems theory) with the input level. Where the effect of this is to reduce the gap, it is called negative feedback, and where the effect of the feedback is to increase the gap, it is called 'positive feedback'.

Kluger & DeNisi (1996) define 'feedback interventions' in the context of human performance as "actions taken by an external agent to provide information regarding some aspects of one's task performance"—in other words all that is required is that there exists some mechanism for evaluating the current level of performance, which is then fed back to the learner. Other definitions of feedback also require that the current level of performance is compared with some desired level (often called the *reference* level), that lies beyond the current level of achievement, and that there exists some mechanism for comparing the two levels, thus establishing the existence of a performance-reference shortfall or a 'gap'. With this definition, if a student who aspires to (say) a score of 75 in a particular assessment (in order to be able to attain certification or licensure), and who has actually scored 50, is told that they need to improve, then this would count as feedback.

That such an arrangement could be described as 'feedback' is rejected by Ramaprasad (1983) who defines feedback as follows:

> "Feedback is information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way" (p4).

In other words, Ramaprasad's definition specifically requires that for feedback to exist, the information about the gap must be used to alter the gap. If the information is not actually used in altering the gap, then there is no feedback.

Which of the definitions of feedback discussed above is used is of course less important than establishing a clarity about the differences between different kinds of feedback. Specifically, feedback functions as

formative assessment (ie guides the form of future learning), only when all the conditions described above are met. Specifically:

1 there must exist a mechanism for evaluating the current level of achievement;
2 a desired level of achievement, beyond the current level of achievement (the *reference* level) must be identified;
3 there exists some mechanism by which to compare the two levels, establishing the *existence* of a 'gap';
4 the learner obtains information about *how* to close the gap;
5 the learner *actually uses this information* in closing the gap.

By reference to relevant research studies, it is argued in the following sections that classroom assessment will be most effective where teachers have appropriate means to elicit information about the learner's current capabilities and direct attention to what it is that learner needs to do to improve (rather than comparing their existing level of achievement with that of other learners). For their part, learners need to be clear about the criteria by which their work will be judged, and are able to monitor and regulate their own progress towards their goal. This is summed up in table 1.

	Telos	Action
Teacher	Eliciting information	Task-involving feedback
Learner	Understanding quality	Self-assessment

Table 1: facets of formative assessment

3.1 Eliciting information

There is considerable evidence that the questioning strategies employed by teachers are often very narrow, are focused primarily on lower-order aspects of capability such as recall and automated skills, and do not allow students sufficient time either to answer, or to elaborate their first attempts at responding.

In most instructional settings, the purpose of questioning is to help the teacher establish what the learner has learned, and specifically, whether the learner has learned what the teacher expected to be learned. The teacher therefore seeks to establish whether there is a discrepancy between the conceptions held by the learner and the conceptions towards which the teacher is aiming. However, *the failure to establish such a discrepancy does not mean that it does not exist*. If the learner answers all the teacher's questions as the teacher expects the questions to be answered by a competent learner, this establishes only that the learner's conceptions *fit* with those of the teacher, within the limitations of the questions asked. It does not establish that the learner's conceptions *matches* those of the teacher. As von Glasersfeld (1987 p13) notes:

> In short, the interviewer is constructing a *model* of the child's notions and operations. Inevitably, that model will be constructed, not out of the child's conceptual elements, but out of the conceptual elements that are the interviewer's own. It is in this context that the epistemological principle of *fit*, rather than *match* is of crucial importance. Just as cognitive organisms can never compare their conceptual organisations of experience with the structure of an independent objective reality, so the interviewer, experimenter, or teacher can never compare the model he or she has constructed of a child's conceptualisations with what actually goes on in the child's head. In the one case as in the other, the best that can be achieved is a model that remains viable within the range of available experience. (p13)

This is brought out very clearly by the following pairs of items from the Third International Mathematics and Science Study (TIMSS).

Item 1: which of the following fractions is the smallest?

Item 2: which of the following fractions in the largest?

The success rate for Israeli middle school students on the first item was 88%, but on the second item, only 46%, with 39% choosing response 'b' (Vinner, 1997 p74-5). An explanation of this is easy to see. The naive strategy of "biggest denominator makes the smallest fraction/smallest denominator makes the biggest fraction" yields the correct response for item 1, but yields response 'b' for item 2. Item 1 is therefore a weaker item than item 2 because students can get the right answer for the wrong reason. A correct answer to item 1 does not indicate a 'fit' with the teacher's conceptions, but merely a 'match' within the limitation of the question.

A second mathematical example is provided by the following pair of linear equations.

$$3a = 24$$

$$a + b = 16$$

Many secondary school students find this item impossible to solve. On detailed questioning, or on giving the students time to talk about their ideas, a student will generally volunteer something like "I keep on getting b is 8 but it can't be because a is 8". Many students establish a belief that letters must stand for *different fixed unknowns*, because, typically, solution of equations is preceded by practice in substitution of numerical values for letters in algebraic expressions where every letter does stand for a different fixed unknown. The important point here is that had the second equation been $a + b = 17$, the students would have solved this item quite happily, and the teacher would have concluded that the students' conceptions matched her own, whereas the meanings attached to these ideas by the students could have been quite different from those the teacher intended.

The role of beliefs in preventing students who have the necessary mathematical understanding from working successfully in mathematics is illustrated clearly in the following example. A twelve-year old girl was working with the diagram shown as figure 2.

Figure 2

The purpose of the diagram was to enable the student to see that by calculating the area of the rectangle in two different ways, the student could see that $6(p+2)$ was equal to $6p+12$. However, the girl steadfastly refused to acknowledge that $6(p+2)$ could be expressed as $6p+12$ unless she was told the value of p. At first, the teacher[4] assumed that the difficulty was based on a misunderstanding of the nature of an algebraic variable as one that could stand for any number. However, after considerable probing it eventually emerged that the student understood the role of variables in algebra perfectly well. The difficulty arose because the student believed that "you have to work out the thing in the brackets first"—a belief established, no doubt, by repeated instruction from the teacher.

Although mathematics classrooms display perhaps the starkest examples of the difference between 'fit' and 'match', similar examples abound in the sciences. For example, after a teacher had taught a class about the molecular structure of water, the students had been asked to draw a beaker of water showing the molecular structure of water. One student, who had shown a section of the beaker with molecules appropriately displayed, was asked what was between the molecules, to which she replied, "Water" (Paul Black, 1998, personal communication). While her diagram showed that she had, apparently, understood the molecular structure of water, deeper questioning showed that her ideas were quite different from those that the teacher had intended.

A necessary pre-condition for effective learning, therefore, is a mechanism for establishing that there are differences between the current level of achievement and the desired level. This requires the availability of a stock of good probes, perhaps in the form of questions, but perhaps also in the form of statements, which, as Dillon (1985) has pointed out, can often be far more productive than questions (which tend to close down classroom discourse).

Having established the existence of a performance-standard shortfall, the likelihood of successful action depends on the learner's reaction to it.

[4] Actually, me.

3.2 Responses to a 'gap'

In their review of the effects of feedback, Kluger and DeNisi (1996) found four responses to the existence of a performance-reference shortfall:

- attempt to reach it (a typical response when the goal is clear, where the individual has a high commitment to achieving the goal and where the individual's belief in eventual success is high);
- abandon the standard completely (particularly common where the individual's belief in eventual success is low, leading to what Carol Dweck (1986) has termed 'learned helplessness');
- change the standard (rather than abandoning it altogether, individuals may lower the standard, especially where they cannot or do not want to abandon it, and conversely, may, if successful, choose to raise the standard);
- deny it exists.

Which of these is the most likely will depend on host of personal, contextual and structural factors. One particularly important issue is the source of the motivation for the desire to close the performance-standard shortfall. Deci and Ryan (1994) have moved beyond the traditional dichotomy between extrinsic and intrinsic motivation to take into account both the locus of motivation and the locus of the value system adopted by the learner.

They propose the term external regulation for those behaviours "that are regulated by contingencies overtly external to the individual" (all quotations from p 6), while introjected regulation "refers to behaviours that are motivated by internal prods and pressures such as self-esteem-relevant contingencies". Identified regulation "results when a behaviour or regulation is adopted by the self as personally important or valuable", although the motivation is extrinsic, while integrated regulation "results from the integration of identified values and regulations into one's coherent sense of self". These four kinds of regulation can therefore be regarded as the result of crossing the locus of the value system with that of the motivation, as shown in table 2.

		value system	
		external	internal
locus of motivation	external	external regulation	identified regulation
	internal	introjected regulation	integrated regulation

Table 2: Classification of behaviour regulation, based on Deci & Ryan (1994)

Within this framework, it can be seen that both internal and external motivation can be effective, but only when associated with internally, as opposed to externally, valued aims. External value systems are much more likely to result in abandoning the attempt to close the performance standard discrepancy, or to deny that it exists at all, while changing the standard would appear to be associated with and internal value system (eg, I was aiming for a grade 'B', but I'm going for a 'C' now).

It is also clear that an important element of the classroom assessment process, which, over the long term, feeds into students' self image and general motivation, as well as influencing the ways in which students react to, and attribute failure, is the kind of feedback they receive on their performance.

3.3 Feedback

Kluger and DeNisi (1996) reviewed over 3000 reports (2500 papers and 500 technical reports) on the effects of feedback, although applying strict quality criteria excluded all but 131 reports involving 12652 participants.

Over the 131 studies the average effect was that feedback did raise performance. The average effect size was around 0.4 (equivalent to raising the achievement of the average student to the 65th percentile). However, this mean figure conceals a great deal of variation in the effect sizes. In 40% of the cases studied, the effect size was negative (albeit small). In other words, the use of feedback had actually lowered performance compared to those given no feedback in two out of every five studies.

It is therefore clear that feedback *per se* is neither 'a good thing' nor 'a bad thing'—what matters is the *quality* of the feedback, and one of the most important features of feedback is whether it is ego involving or task-involving—in other words, whether it directs attention to the self, or to the task.

This is well illustrated by the work of Ruth Butler, who, in a series of studies, has shown the importance of making sure that feedback is task-involving rather than ego-involving. In one study, (Butler, 1988) 48 eleven-year old Israeli students were selected from the upper and lower quartiles of attainment from 12 classes in 4 schools and worked in pairs over three sessions on two tasks (one testing convergent thinking and the other, divergent).

After each session, each student was given written feedback on the work they had done in the session in one of three forms:

A individualised comments on the extent of the match of their work with the assessment criteria that had been explained to each class at the beginning of the experiment;

B grades, based on the quality of their work in the previous session;

C both grades and comments.

Students given comments showed a 30% increase in scores over the course of the experiment, and the interest of all students in the work was high. Students given only grades showed no overall improvement in their scores, and the interest of those who had high scores was positive, while those who had received low scores show low interest in the work. Perhaps most surprisingly, the students given both grades and comments performed similarly to those given grades alone—no overall improvement in scores, and interest strongly correlated with scores—and the researchers themselves describe how students given both grades and comments ignored the comments, and spent their time comparing their grades with those achieved by their peers.

An explanation of this effect is suggested by another study by Butler (1987) in which 200 grade 5 and grade 6 Israeli school children were given one of four kinds of feedback (comments, grades, praise, no feedback) on their performance in divergent thinking tasks. Although there were no significant differences between the groups receiving the different kinds of feedback in their pre-test scores, the students receiving comments scored one standard deviation higher than the other groups on the post-test (there were no significant differences between the other three groups). Furthermore, students given grades and praise scored far higher than the 'comments' or the 'no feedback' groups on measures of ego-involvement, while those given comments scored higher than the other three groups on measures of task-involvement. In other words, grades and praise had no effect on performance, and served only to increase ego-involvement.

The clear message from this, and many other studies reviewed by Black and Wiliam, is that for feedback to be effective, it must be focused on what the individual student needs to do to improve (ie it must be *task-involving*) rather than focusing attention on to the learner and her or his self-esteem (ie *ego-involving*).

However, what is not yet clear is whether the task-involving feedback should be focused at the general level (eg "I think it would be a good idea if you gave some more thought to the middle section of your essay), or related to specific features of the task (eg your use of tense in the third and fourth paragraphs is inconsistent). While at first sight it might seem clear that detailed, task-specific feedback would be better, such feedback might 'take the learning away from the student', and having to find out what needs to be improved might be an important element in coming to understand what counts as a good piece of work (see 'Sharing criteria with learners' below). Certainly there is evidence that too much feedback can be counterproductive. For example, in a study of 64 grade 3 students, Day and Cordón (1993) found that those students who were given what they called a 'scaffolded' response when stuck—ie the minimum help necessary to allow them to continue with their work—did substantially better than those given a complete solution to the task, and then given a second task on which to work. The students given the 'scaffolded' approach also did better on other tasks, whether closely related to the tasks on which they had worked or not. Perhaps most importantly, the final achievement levels of the students given the scaffolded approach varied little and were unrelated to measures of general ability and prior achievement, whereas the final achievement levels of those given complete solutions when 'stuck' were strongly related to measures of both general ability and prior achievement.

3.4 Ensuring learners understand quality

There is considerable evidence that many students in classrooms do not understand what it is that teachers value in their work—perhaps the starkest illustration of this is the old definition of project work as being "four weeks on the cover and two on the contents".

The importance of sharing criteria with learners is shown by an important experiment by Frederiksen and White (1997), carried out in 12 classes of 30 students each in two schools in a socio-economically disadvantaged area in the United States. For a proportion of the lesson time available each week, each class was divided into two halves. One half of each class used some of the class time for general discussion of the module, while the other half of the class spent the same amount of time on a discussion that was structured to promote reflective assessment (involving, for example, peer assessment of presentations to the class and self-assessment). Apart from these sessions when the class was divided, all other science lessons were taken together, in the same classroom, with the same teacher.

In order to establish whether the learning outcomes were similar for different kinds of students, before the experiment began all the students involved had taken a test of basic skills. The students' learning at the end of the teaching experiment was assessed in three ways: the mean score achieved on all the projects undertaken during the experiment, the score obtained on two chosen projects which the students had carried out independently, and a score on a physics concepts test. The pattern of results on all three measures of outcome were broadly similar. In general, the students who had experienced the 'reflective assessment' achieved higher scores across the board than the 'control' groups. Perhaps more importantly, these gains were more marked for the students who gained the lowest marks on the basic skills test.

For example, on the mean project scores, the experimental students getting the highest marks on the basic skills test out performed their counterparts in the control group by one standard deviation, but the lowest-scoring one-third of the experimental students (as measured by the basic skills test) outperformed their counterparts in the control group by *three standard deviations*.

For successful learning to take place, therefore, it is seems highly desirable for the learner to come to understand the criteria against which her or his work will be assessed. Indeed Sadler (1989) regards it as an essential pre-requisite for learning that *the learner comes to understand the goals towards which she is aiming*. In this era of criterion-referenced hyperspecification (Popham, 1994), it has become commonplace (see for example, Klenowski, 1995) to require that the criteria for success should be explicit, predetermined and general (rather than implicit, developed after the work has been completed, and tailored to the particularities of the specific piece of work). But this brings us back to the argument in section 2.3—such criteria cannot convey their meanings unambiguously. For example, if a teacher tells a student that she needs to "be more systematic" in (say) her scientific explorations, 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 teacher believes the advice she is giving is helpful, but that is because the teacher already knows what it means to be systematic. However if the learner understood what "being systematic" meant, she would probably have been able to be more systematic in the first place. This is exactly the same problem with criterion-referenced assessment encountered in section 2.2. The solution is to work towards *bringing learners into the same community of practice of which the teachers are already members*. As before, maxims can be used, not as definitions, but as starting points for negotiating meaning so that the learners come to share the implicit standards of quality already possessed by the teacher.

3.5 Self-assessment

As well as understanding the goals against which they will be assessed, it is also clear that there are significant benefits if students are encouraged to monitor their own progress towards these goals. For example, a study by Fontana and Fernandes (1994) compared the progress of students (aged from 8 to 14) taught by teachers undertaking the same amount of inservice training (one evening session per week for 20 weeks). Approximately half the pupils were taught by teachers participating in general inservice education while the teachers of the other concentrated on using student self-assessment in the classroom . There was no difference between the two groups of teachers in their qualifications, experience, the amount of time spent on teaching, nor in the classroom materials used). However, the progress made by the teachers using self-assessment was *twice* that of the other students.

The evidence outlined above, and the more substantial body of evidence presented in Black and Wiliam, therefore provides clear evidence that standards of achievement will be raised considerably if teachers can elicit appropriate information about students' conceptions, can feed this information back to students

in a way that focuses on what it is the student needs to do in order to improve (rather on the student's self-esteem), can inculcate in the students an implicit understanding of what counts as quality in their work, and can equip the students with the ability to monitor their own progress towards their goals.

What matters, therefore, in formative assessment, is that the assessment yields information which provides a recipe for future action, either by the learner or the teacher. In particular, it does not matter if two different teachers have different interpretations of the results of an assessment designed for a formative purpose if for both teachers, the different interpretations lead to successful learning for the learner. Adopting the distinction between evidential and consequential bases for result and interpretation and use from Messick (1980), it seems reasonable to conclude that summative assessments are validated by their *meanings*, while formative assessments are validated by their *consequences* (see Wiliam & Black, 1996 for a fuller discussion of this point).

4. Integrating Formative and Summative Assessment

The arguments presented above indicate that high-quality educational provision requires that teachers are involved in both summative and formative assessment. Some authors (eg Torrance, 1993) have argued that the formative and summative functions of assessment are so different that the same assessment system cannot fulfil both functions.

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. The result is always that the formative assessment system is 'driven out' by that for summative assessment. On the assumption that this is true in practice (whether or not it is logically necessary), then there are only three possibilities:

- remove teachers' responsibility for summative assessment
- remove teachers' responsibility for formative assessment
- find ways of ameliorating the tension between summative and formative functions of assessment.

The arguments in Section 2 establish the undesirability of the first of these and those of Section 3 establish the same for the second. 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.

Of course, this is a vast undertaking, and well beyond the scope of this, or any other single paper. The remainder of this paper is therefore intended to suggest some theoretical foundations that would allow the exploration of possibilities for mitigating, if not completely reconciling, the tension between formative and summative assessment.

To sum up the argument so far, traditional forms of assessment have started with the idea that the primary purpose of educational assessment was selecting and certifying the achievement of individuals and we have 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 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. This necessitates refining the distinction between formative and summative functions of assessment

4.1 The functions of assessment

For the remainder 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.

Of course other authors have used these terms in different ways. In 1967 in an AERA monograph on evaluation, it was Michael Scriven who first 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, as noted above, 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.

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.

4.2 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 noted earlier in which a certain amount of knowledge or capability is believed to exist within the individual being assessed, and evidence of which is generated more or less reliably by the assessment. 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, to return to the test of number bonds described earlier, if a student answered questions on all twenty number bonds correctly, we know that, on at least one occasion, that she or he could answer correctly questions on those twenty bonds, 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.

As noted earlier, the typical 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 number bonds 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 bonds matter (at least for the moment), with the social consequence that only those twenty bonds get learned.

In terms of Messick's model of validity argument (figure 1) 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 number bonds 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 (Goodhart's law again), so that in an extreme case (such as our number bonds 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.

4.3 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.

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"! Now as noted above, if learners come to share the construct of quality held by the teacher, they too would be able to say that a particular piece of work merited (say) a 'D'. What we cannot expect the learners to be able to do, and which therefore must be left to the teacher, is to be able to identify what would need to be changed for this (say) 'D' piece of work to be worth a 'C'? In other words, while both teachers and learners must share a *construct of quality*, it is necessary for a teacher also to have an *anatomy of quality*. They must be able to identify what would be the most fruitful next step on the road to improvement.

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? [as mentioned in section 2.2]

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, such items 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 and 407 statements of attainment respectively, 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 diagnostic 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, like diagnostic and formative 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.

4.4 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*.

5 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.

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 *pro*spective. They must contain within themselves a recipe for future action, and

their validity consists 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.

6 REFERENCES

Allal, L. & Pelgrims Ducrey, G. (2000). Assessment of—or in—the zone of proximal development. *Learning and Instruction,* **10**(2), 137-152.

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

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. (1998a). Assessment and classroom learning. *Assessment in Education: Principles Policy and Practice,* **5**(1), 7-73.

Black, P. J. & Wiliam, D. (1998b). *Inside the black box: raising standards through classroom assessment*. London, UK: King's College London School of Education.

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**(5), 631-648.

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.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: effects of different feedback conditions on motivational perceptions, interest and performance. *Journal of Educational Psychology,* **79**(4), 474-482.

Butler, R. (1988). Enhancing and undermining Intrinsic motivation; the effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology,* **58**, 1-14.

Claxton, G. (1995). What kind of learning does self-assessment drive? Developing a 'nose' for quality; comments on Klenowski. *Assessment in Education,* **2**(3), pp. 339-343.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin,* **52**(4), 281-302.

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.

Day, J. D. & Cordon, L. A. (1993). Static and dynamic measures of ability: an experiemntal comparison. *Journal of Educational Psychology,* **85**(1), 76-82.

Deci, E. L. & Ryan, R. M. (1994). Promoting self-determined education. *Scandinavian Journal of Educational Research,* **38**(1), 3-14.

Department for Education & Welsh Office (1995). *English in the National Curriculum*. London, UK: Her Majesty's Stationery Office.

Dillon, J. T. (1985). Using questions to foil discussion. *Teaching and Teacher Education,* **1**(2), 109-121.

Doyle, W. (1983). Academic work. *Review of Educational Research,* **53**(2), 159-199.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist (Special Issue: Psychological science and education),* **41**(10), 1040-1048.
Fontana, D. & Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portugese primary school pupils. *British Journal of Educational Psychology,* **64**, 407-417.
Foucault, M. (1977). *Discipline and punish* (Sheridan-Smith, A M, Trans.). Harmondsworth, UK: Penguin.
Frederiksen, J. R. & White, B. J. (1997). Reflective assessment of students' research within an inquiry-based middle school science curriculum. In Proceedings of *Annual meeting of the AERA* conference, vol . Chicago, IL.
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.
Hanson, F. A. (1993). *Testing testing: social consequences of the examined life*. Berkeley, CA: University of California Press.
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.
Hill, C. & Parry, K. (1994). Models of literacy: the nature of reading tests. In C. Hill & K. Parry (Eds.), *From testing to assessment: English as an international language* (pp. 7-34). Harlow, UK: Longman.
Kellner, P. (1997). Hit-and-miss affair. *Times Education Supplement*, 23.
Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and cearning contexts of Australia and England. *Assessment in Education: principles, policy and practice,* **2**(2), 145-163.
Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention Theory. *Psychological Bulletin,* **119**(2), 254-284.
Lave, J. & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.
Linn, R. L. (1994) *Assessment-based reform: challenges to educational measurement.* Paper presented at Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service.
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.
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.
Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (pp. 13-103). Washington, DC: American Council on Education/Macmillan.
Morgan, C. (1998). *Writing mathematically: the discourse of investigation*. London, UK: Falmer Press.
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*. Chicago, IL: University of Chicago Press.
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.
Popham, W. J. (1994, April) *The stultifying effects of criterion-referenced hyperspecification: a postcursive quality control remedy.* Paper presented at Symposium on Criterion-referenced clarity at the annual meeting of the American Educational Research Association held at New Orleans, LA. Los Angeles, CA: University of California Los Angeles.
Ramaprasad, A. (1983). On the definition of feedback. *Behavioural Science,* **28**(1), 4-13.
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.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: new tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments : alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston, MA: Kluwer Academic Publishers.
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.
Searle, J. R. (1995). *The construction of social reality*. London, UK: Allen Lane, The Penguin Press.
Shavelson, R. J.; Baxter, G. P. & Pine, J. (1992). Performance assessments: political rhetoric and measurement reality. *Educational Researcher,* **21**(4), 22-27.
Smith, M. L. (1991). Meanings of test preparation. *American Educational Research Journal,* **28**(3), 521-542.
Torrance, H. (1993). Formative assessment: some theoretical problems and empirical questions. *Cambridge Journal of Education,* **23**(3), 333-343.
Vinner, S. (1997). From intuition to inhibition—mathematics, education and other endangered species. In E. Pehkonen (Ed.) Proceedings of *21st Conference of the International Group for the Psychology of Mathematics Education* conference, vol 1 (pp. 63-78). Lahti, Finland: University of Helsinki Lahti Research and Training Centre.
von Glasersfeld, E. (1987). Learning as a constructive activity. In C. Janvier (Ed.) *Problems of representation in the teaching and learning of mathematics* Hillsdale, NJ: Lawrence Erlbaum Associates.
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. & 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.
Wiliam, D. (1992). Some technical issues in assessment: a user's guide. *British Journal for Curriculum and Assessment,* **2**(3), 11-20.
Wiliam, D. (1993). Validity, dependability and reliability in national curriculum assessment. *The Curriculum Journal,* **4**(3), 335-350.
Wiliam, D. (1994). Assessing authentic tasks: alternatives to mark-schemes. *Nordic Studies in Mathematics Education,* **2**(1), 48-68.
Wiliam, D. (1994). Assessing authentic tasks: alternatives to mark-schemes. *Nordic Studies in Mathematics Education,* **2**(1), 48-68.
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. (1996a). National curriculum assessments and programmes of study: validity and impact. *British Educational Research Journal,* **22**(1), 129-141.
Wiliam, D. (1996b). Standards in examinations: a matter of trust? *The Curriculum Journal,* **7**(3), 293-306.

Address for correspondence: Dylan Wiliam, School of Education, King's College London, Franklin-Wilkins Building, Stamford Street, London SE1 9NN, England.
Tel: +44 20 7848 3153; Fax: +44 20 7848 3182; Email: dylan.wiliam@kcl.ac.uk.