

# **An overview of the relationship between assessment and the curriculum\***

**Dylan Wiliam  
King's College London School of Education**

## **Introduction**

For most of this century, and in most countries, the most common method for certifying achievement in schools, whether for purposes of social accountability, or for providing information to aid decisions on the futures of individuals, has been by the administration of an assessment instrument that is devised, and is scored, externally.

These external 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, it has become increasingly clear over the past twenty years that the contents of standardised tests and examinations are not a random sample from the domain of interests. In particular, these 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 is one of the major reasons underlying the shift in interest towards 'authentic' or 'performance' assessment (Resnick & Resnick, 1992)—assessments that measure valued performance like writing essays, undertaking scientific experiments, solving complex mathematical problems and so on, directly, rather than through the use of proxies like multiple-choice or short-answer tests.

In high-stakes settings, performance on standardised tests can not be relied upon to be generalizable to more authentic tasks. 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).

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 individual authentic tasks, the variability of tasks is a significant issue. In other words, the score that a student gets on a specific task depends partly on how good the student is, but also on whether that particular task suited the student's strengths and weaknesses. If we use only a small number of tasks, then the overall score achieved by students will depend to a significant extent on whether the particular tasks they were asked to do suited them—in other words, we are assessing how lucky they are as much as how competent they are in the domain being assessed. Using authentic tasks improves validity, in

---

\* From *Curriculum and assessment*, edited by David Scott (pp. 165-181), Greenwich, CT: JAI Press in 2001.

that they tell us about students' performance on important aspects of the domain that are generally neglected in multiple-choice and short-answer tests, but reliability is generally weakened, in that the results of authentic tasks taking the same amount of time as multiple-choice tests are generally less reliable.

This can be illustrated by drawing an analogy with stage lighting. For a given power of illumination, we can either focus this as a spotlight or as a floodlight. The spotlight brings real clarity to a small part of the stage, but the rest of the stage is in darkness. This is analogous to a highly-reliable multiple-choice test, in which the scores on the actual matter tested are highly reliable, but we know nothing about the other aspects of the domain that were not tested. A floodlight, on the other hand illuminates the whole stage. We may not be able to make quite such accurate distinctions in the small part of the domain assessed by the multiple-choice test, but what we can say about the other areas will be *more* accurate.

The work of Shavelson, Baxter, & Pine, (1992) shows that we don't get adequately reliable results even in subjects like mathematics and science unless we use at least six tasks, and in other subjects, where students' liking of the task may be more important, we may need ten or more.

Since it is hard to envisage many worthwhile authentic tasks that could be completed in less than an hour or two, the amount of assessment time that is needed for the reliable 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. This also has the effect of integrating learning and assessment, which is taken up in more detail below. 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 in most countries 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 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 marking 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. And 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 German commentator once remarked: 'Why rely on an out-of-focus snapshot taken by a total stranger?'

The arguments outlined above suggest that high-quality educational provision *requires* that teachers are involved in the summative assessment of their students. However, it is also clear that high quality educational provision requires effective formative assessment as well (see Black, this volume). Are the formative and summative functions of assessment compatible? Some authors (eg Torrance, 1993) have argued that formative and summative assessment are so different that the same assessment system cannot fulfil both functions. Maintaining dual assessment systems would appear to be quite simply beyond the capabilities of the majority of teachers, with the formative assessment system being driven out by that for summative assessment. If this is true in practice (whether or not it is logically necessary), then there are only three possibilities:

- teachers are not involved in the summative assessment of their students
- teachers are not involved in the formative assessment of their students
- we find ways of ameliorating the tension between summative and formative functions of assessment.

In view of the foregoing arguments, I consider the consequences of the first two of these possibilities to be unacceptable, and therefore I would argue that if we are to try to create high-quality educational provision, ways must be found of mitigating the tension between formative and summative functions of assessment.

Of course, this is a vast undertaking, and well beyond the scope of this, or any other single article. The remainder of this chapter is therefore intended simply 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.

### **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 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 & Meehl noted over forty years ago, "One does not validate a test, but only a principle for making inferences" (Cronbach & Meehl, 1955 p297).

More recently, it has become more generally accepted that it is also important to consider the *consequences* of the use of assessments as well as the validity of inferences based on assessment outcomes. Some authors have argued that a concern with consequences, while important, go beyond the concerns of validity—George Madaus for example uses the term *impact* (Madaus, 1988). 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 the validation of assessments

The upper row of Messick's table relates to traditional conceptions of validity, while the lower row relates to the *consequences* of assessment interpretation and use. One of the most important 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. The assessments do not just represent the values associated with the domain, but actually serve to define them—what gets assessed tells us what the subject is 'really' about, and teachers and students act accordingly.

The use of Messick's framework can be illustrated by considering whether a student's competence in speaking and listening in the mother tongue should be assessed in an assessment of their overall competence in the language. Each of the following sets of arguments relates to one of the cells in figure 1.

- A Many authors have argued that an assessment of English that ignores speaking and listening skills does not adequately represent the domain of 'English'. This is an argument about the evidential basis of result interpretation (such an assessment would be said to under-represent the construct of 'English').
- B There might also be empirical evidence that omitting speaking and listening from an assessment of English reduces the correlation with other accepted assessments of the same domain (concurrent validity) or with some predicted outcome, such as advanced study (predictive validity). Either of these would be arguments about the evidential basis of result use.
- C It could certainly be argued that leaving out speaking and listening would send the message that such aspects of English are less important, thus distorting the values associated with the domain (consequential basis of result interpretation).
- D Finally, it could be argued that unless such aspects of speaking and listening were incorporated into the assessment, then teachers would not teach, or would place less emphasis on, these aspects (consequential basis of result use).

Messick's model presents a useful framework for the structuring of validity arguments, but it provides little guidance 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.

#### *Referents in 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 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*. 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.

#### *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, as Angoff (1974) pointed out, any criterion-referenced assessment is under-pinned by a set of norm-referenced assumptions, because the assessments are used in social settings and for social purposes. 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 have little or none.

Furthermore, no matter how precisely the criteria are drawn, it is clear that some judgement must be used—even in mathematics—in deciding whether a particular item or task performance does yield evidence that the criterion has been satisfied (Wiliam, 1993).

Even if it were possible to define performance domains unambiguously, it is by no means clear that this would be desirable (Mabry, 1999). 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. Put crudely, the more precisely we specify what we want, the more likely we are to get it, but the less likely it is to mean anything.

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 and it seems likely that the contrast between norm and criterion-referenced assessment represents the concerns of, and the kinds of assessments developed by, specialists in educational and psychological measurement. Beyond these narrow concerns there are a range of assessment events and assessment practices that are typified by the traditions of school examinations in European countries, and by the day-to-day practices of teachers all over the world. These practices rely on authentic rather than indirect assessment of performance, and are routinely interpreted in ways that are not faithfully or usefully described by the contrast between norm and criterion-referenced assessment.

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. An alternative approach is, instead of building

theoretical models and then trying to apply them to assessment practices, we try to theorise what is actually being done. After all, 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.

### **Construct-referenced assessment**

The model of 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 (the General Certificate of Secondary Education or GCSE). Until the government's recent change in national examinations, which required all GCSEs to have an externally-assessed component, the GCSE grade for the vast majority of students in England and Wales was determined not by performance on an examination, but entirely on the basis of a portfolio of work, prepared by the student, and assessed by her or his teacher. 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 no attempt is made to prescribe learning outcomes. In that it is defined at all, it is defined simply as the consensus of the teachers making the assessments. The assessment is not objective, in the sense that there are no objective criteria for a student to satisfy, but the experience in England is that it can be made reliable. To put it crudely, it is not necessary for the 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 practitioners (Lave & Wenger, 1991), I have proposed elsewhere that such assessments are best described as 'construct-referenced' (William, 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 regulations of the University of London regulations 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 times the letter 'e' occurs in this book is, currently, I am sure, not known to anyone, so simply counting occurrences of the letter 'e' this book would generate a new fact, but there is surely not a university in the world that would consider it 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. The consistency of such assessments depends on what Polanyi (1958) called *connoisseurship*, but perhaps might be more useful regarded as the membership of a community of practice (Lave & Wenger, 1991).

The touchstone for distinguishing between criterion- and construct-referenced assessment is the relationship between the written descriptions (if they exist at all) and the domains. Where written statements collectively *define* the level of performance required (or more precisely where they define the justifiable inferences), then the assessment is criterion-

referenced. However, where such statements merely *exemplify* the kinds of inferences that are warranted, then the assessment is, to an extent at least, construct-referenced.

### **How to do things with assessments: illocutionary speech acts and communities of practice**

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. In contrast, perlocutionary speech acts are speech acts *about* what has, is or will be. For example, the verdict of a jury in a trial is an illocutionary speech act—it does what it says, since the defendant becomes innocent or guilty simply by virtue of the announcement of the verdict. Once a jury has declared someone guilty, they *are* guilty, whether or not they really committed the act of which they are accused, until that verdict is set aside by another (illocutionary) speech act.

Another example of an illocutionary speech act is the wedding ceremony, where the speech act of one person (the person conducting the ceremony saying “I now pronounce you husband and wife”) actually does what it says, creating what John Searle calls ‘social facts’ (Searle, 1995).

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 summative assessments are treated as if they were perlocutionary speech acts, whereas they are perhaps more usefully 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 awarded as the result of a speech act of a single person (ie the examiner where there is just one, or the chair of the panel where there are more than one). 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 do now that 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 assessment, duly authorised bodies 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. Until recently, the same was true of the award of a high-school diploma in the USA.

Now the *existence* of a community of practice is no evidence of its legitimacy. There is an inevitable tendency for such communities of practice to reproduce themselves by admitting only 'people like us'. But the *authority* of such communities of practice (as opposed to *power*) will depend on the trust that individuals beyond that community are prepared to place in its judgements. In order to maintain this trust communities will have to show that their procedures for making judgements are fair, appropriate and defensible (ie that they are *valid*), even if they cannot be made totally transparent, and the paper by Anthony Nitko (this volume) provides a framework within which this can be accomplished.

The foregoing theoretical analysis creates, I believe, a framework for the validation of teachers' summative assessments of their students. 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 is a statement that the performance is adequate to inaugurate the student into a community of practice.

The advantage of such a system of summative assessment is that the evidence-base for the assessment is grounded in the learning environment, so that it can also support formative assessment.

### **Formative assessment**

Strictly speaking, there is no such thing as a formative assessment. The formative-summative distinction applies not to the assessment itself, but to the use to which the information arising from the assessment is put. The same assessment can serve both formative and summative functions, although in general, the assessment will have been designed so as to emphasise one of the functions.

As noted by Black (this volume) formative assessment can be thought of "as encompassing all those activities undertaken by teachers and/or by their students which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (Black and Wiliam, 1998a).

Although perhaps somewhat simplistic, it is useful to break this general idea into three (reasonably distinct) phases: the elicitation of evidence regarding achievement, the interpretation of that evidence, followed by appropriate action.

The evidence of achievement provides an indication of the actual level of performance, which is then interpreted relative to some desired or 'reference' level of performance. Some action is then taken to reduce the gap between the actual and the 'reference' level. The important thing here—indeed some would argue the defining feature of formative assessment—is that the information arising from the comparison between the actual and desired levels *must* be used in closing the gap. If, for example, the teacher gives feedback to the student indicating what needs to be done next, this will not be formative unless the learner can understand *and act* on that information. An essential pre-requisite for assessment to serve a formative function is therefore that the learner comes to understand the goals towards which she is aiming (Sadler, 1989). If the teacher tells the student that she needs to

“be more systematic” in her mathematical investigations, that is not feedback unless the learner understands what “being systematic” means—otherwise this is no more helpful than telling an unsuccessful comedian to “be funnier”. The difficulty with this is that if the learner understood what “being systematic” meant, she would probably have been able to be more systematic in the first place. The teacher believes the advice she is giving is helpful, but that is because the teacher already knows what it means to be systematic. This is exactly the same issue we encountered in the discussion of criterion-referenced assessment above, and why I believe, in contrast to Klenowski (1995), that learning goals can never be made explicit. The words used—whether as criteria or for feedback—do not carry an unambiguous meaning, and require the application of implicit knowledge (Claxton, 1995).

Now this should not be taken to mean that ‘guidelines’ or ‘criteria’ should not be used in helping learners come to understand the goals the teacher has in mind. These criteria can be extraordinarily helpful in helping learners begin to understand what is required of them. But it is a fundamental error to assume that these statements, however carefully worded, have the same meaning for learners as they do for teachers. Such statements can provide a basis for negotiating the meaning, but ultimately, the learners will only come to understand the statements by seeing them exemplified in the form of actual pieces of students’ work.

This notion of ‘understanding the standard’ is the theme that unifies summative and formative functions of assessment. Summative assessment requires that teachers become members of a community of practice, while formative assessment requires that the learners themselves become members of the same community of practice. As the paper by Broadfoot et al (this volume) makes clear, as well as understanding the cognitive aims of the community of practice, becoming a full participant also requires understanding how the classroom ‘works’, with the students “given a central role in the management of their own learning, but are also given the knowledge and skills to discharge their responsibilities” (Simpson, this volume).

This process of becoming *attuned* to the *constraints* and *affordances* (Gibson, 1979) of the classroom is an essential part of being an effective learner. Whether success in one particular classroom is effective beyond that classroom depends on the extent to which the constraints and affordances of that classroom are available in other settings. Boaler (1997) provides a stark example of students who were highly successful in one particular community of practice, but because the constraint and affordances to which they had become attuned were not present in their examinations, their performance was considerably weakened.

For the teacher’s part, however, as both Black, and Simpson (this volume) point out, it is not enough just to ‘understand the standard’. Where a learner understands the standard, and is able to assess her or his own performance, they can become aware of the ‘gap’ between current and desired achievement. What they lack, however, is any clear idea of how to go about closing the gap. They know *that* they need to improve, but they are unlikely to have any clear idea of *how* to improve (for if they did, they would be able to reach the desired level). An essential role for the teacher in formative assessment is therefore to *analyse* the gap between present and desired performance, and be able to break this down into small, comprehensible steps that can be communicated to the learner (recall the teacher quoted by Simpson who realised that he had, in the past been telling his pupils that they ‘must work harder at problem solving’). Put crudely, summative assessment requires teachers to understand the standard, while formative assessment requires learners to understand the standard, and for teachers to understand the standard and the ‘gap’.

The summative and formative functions of assessment are further distinguished by how they are validated. With summative assessments any unfortunate consequences tend to be justified by the need to establish consistency of meanings of the results across different contexts and assessors. With formative assessment, any lack of shared meanings across different contexts is irrelevant—all that matters is that they lead to successful action in support of learning. In a very real sense, therefore, summative assessments are validated by their meanings and formative assessments by their consequences.

The foregoing theoretical analysis provides a basis for distinguishing between formative and summative functions of assessment, but does not address the issue raised earlier in this paper

and by Val Klenowski (this volume) of the tension between the formative and summative functions of assessment.

As Klenowski shows, in the context of portfolio assessment, the requirements of the summative function for a portfolio to contain particular elements results in a situation in which the formative function is weakened.

Of course, the formative and summative functions of assessment will always be in tension, but the identification of three phases of the assessment cycle above (elicitation, interpretation, action) suggests some ways in which the tension can be mitigated somewhat (for a fuller version of this argument, see Wiliam, 1999).

When evidence is being elicited, 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 they 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. In the case of Klenowski's example of teacher training, we might insist that the portfolio includes statements about the procedures used by the student in evaluating their own practice rather than insisting on actual examples of the evaluations.

Once evidence is elicited, it must be interpreted differently for different purposes, and it is important to note that once the data has been interpreted for one purpose, it cannot easily serve another. For formative purposes, the focus will 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 very difficult assignments can be especially significant, because they can point clearly to learning needs that were not previously clear. However, the results of these difficult assignments should not count against the learner for summative purposes—what goes into the portfolio, for example, must be only a selection from all possible work, and may even be re-drafted or re-worked before it is included. The relationship between the summative and the formative assessment is not the aggregation of the latter into the former, but rather the result of a re-assessment, for a different purpose, of the original evidence.

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

There is no doubt that, for most of the school year, the formative function should predominate:

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

As part of their day-to-day work, teachers will be collecting evidence about their students, and, for most of the year, this will be interpreted with a view to gauging the future learning needs of the students, and helping the students to understand what it would mean to be a member of the community of practice. In such a system “assessment is seen as continuous, concerned with the creation of a flow of contemporary information on pupil progress which will genuinely inform the teaching and learning processes” (Simpson, this volume).

However, at intervals (perhaps only as often as once each year) the original evidence of attainment can be re-visited and re-interpreted holistically, to provide a construct-referenced

assessment that is synoptic of each student's achievement—an indication of the extent to which they have become full members of the community of practice.

## References

- Angoff, W. H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review*, 92 (Summer), 2-5, 21.
- 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.
- Boaler, J. (1997). *Experiencing school mathematics: teaching styles, sex and setting*. Buckingham, UK: Open University Press.
- 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.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. London, UK: Houghton Mifflin.
- Hanson, F. A. (1993). *Testing testing: social consequences of the examined life*. Berkeley, CA: University of California Press.
- 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.
- Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. *Assessment in Education*, 2(2), pp. 145-163.
- Lave, J. & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.
- 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.
- Mabry, L. (1999). Writing to the rubric: lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan*, 80(9), 673-679.
- 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. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (pp. 13-103). Washington, DC: American Council on Education/Macmillan.
- 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.
- 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.
- 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.
- 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. (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. (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.
- Wiliam, D. (1999, August) "There is no alternative": mitigating the tension between formative and summative functions of assessment. Paper presented the 8th biennial meeting of the European Association for Research on Learning and Instruction held at Gothenburg, Sweden. London: King's College London School of Education.