

The validity of teachers' assessments¹

Dylan Wiliam
King's College London School of Education

Introduction

While many authors have argued that formative assessment—that is in-class assessment of students by teachers in order to guide future learning—is an essential feature of effective pedagogy, empirical evidence for its utility has, in the past, been rather difficult to locate. A recent review of 250 studies in this area (Black & Wiliam, 1998a) concluded that there was overwhelming evidence that effective formative assessment contributed to learning. 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 like New Zealand, Germany, the United States 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 & Wiliam, 1998b pp 5-6), and therefore considerable improvements in learning can be achieved by effective implementation of formative assessment.

There are also strong reasons why teachers should also be involved in the summative assessment of their students, for the purposes of certification, selection and placement. External assessments, typically by 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). In high-stakes settings, performance on standardised tests can no longer 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 (Linn & Baker, 1996) has shown that in the assessment of authentic tasks, there is a considerable degree of task variability. In other words, the

¹ Paper presented to Working Group 6 (Research on the Psychology of Mathematics Teacher Development) of the 22nd annual conference of the International Group for the Psychology of Mathematics Education; Stellenbosch, South Africa, July 1998.

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 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. 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 has remarked: ‘Why rely on an out-of-focus snapshot taken by a total stranger?’

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 formative and summative assessment are so different that the same assessment system cannot fulfil both functions. Most countries have found that maintaining dual assessment systems is 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

- 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.

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 assessment.

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 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 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 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 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 investigative work in mathematics is not formally assessed, this is often interpreted as an

implicit statement that such aspects of mathematics 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 open-ended and investigative work into 'high-stakes' assessments of mathematics 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 an assessment of mathematics that ignores open-ended and investigative work does not adequately represent the domain of mathematics. This is an argument about the evidential basis of result interpretation (such an assessment would be said to under-represent the construct of 'Mathematics').
- B It might also be argued that leaving out such work 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 open-ended and investigative work in mathematics would send the message that such aspects of mathematics are not 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 mathematics were incorporated into the assessment, then teachers would not teach, or place less emphasis on, these aspects (consequential basis of result use).

Messick's four-facet 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 'referents' 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* 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.

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 do not.

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. 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. 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, typified by the traditions of school examinations in European countries, and by the day-to-day practices of teachers, 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.

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.

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. 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 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 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 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 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).

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 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 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 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 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.

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.

Formative assessment involves elicitation of evidence regarding achievement, interpreting that evidence, and then taking 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 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, that is not formative unless the learner can understand *and act* on that information. An essential prerequisite 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, 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. The first requires that teachers become members of a community of practice, while the second requires that the learners become members of the same community of practice. The two functions are *distinguished* by how they are validated. With summative assessments any unfortunate consequences are 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 also suggests how the tension between formative and summative functions of assessment can be mitigated. It is clear that once evidence has been interpreted for a summative purpose, it cannot then serve a formative function. However, provided a clear separation is made between the elicitation of the evidence and its interpretation, it seems likely that the same evidence *can* be interpreted in different ways to serve the different functions.

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. However, at intervals (perhaps only as often as once

each year) the original evidence of attainment could then 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.
- 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.
- 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.
- 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, 4.4.94 - 8.4.94) *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.

Address for correspondence: Dylan Wiliam, Dean and Head of School, School of Education, King's College London, Cornwall House, Waterloo Road, London SE1 8WA, England. Email: dylan.wiliam@kcl.ac.uk; Fax: +44 171 872 3153; Tel: +44 171 872 3182.