Assessing authentic tasks: alternatives to mark-schemes

Dylan Wiliam

The kinds of authentic tasks that have been used in national assessments in England and Wales over the last thirty years – typically open-ended, 'pure' investigative tasks – are described, and the marking schemes used for their assessment are classified as either task-specific or generic. Generic schemes are further classified according to whether the 'degree of difficulty' of the task or the 'extent of progress' through the task is given most emphasis. A view of validation is presented that requires consideration of the value implications and social consequences of implementing assessment procedures, and it is argued that both task-specific and generic schemes will have the effect of stereotyping student approaches to these tasks. An alternative paradigm to norm-referenced and criterion-referenced interpretations of assessments, entitled 'construct-referenced' assessment, is proposed as being more consistent with the rationale behind such authentic assessments. Suggestions for the implementation of such a system are made and indices derived from signal-detection theory are suggested as appropriate measures for the evaluation of the accuracy of such assessments.

1 Introduction

For most of this century, there have been two levels of national examinations in England and Wales: one intended for 16-year-old students, which has come to serve as a school-leaving examination, and one taken at age 18, for university entrance.

These examinations and other 'high-stakes' assessments in mathematics have always involved a preponderance of constructed-response questions. Indeed, at university entrance level it has been common to find a three-hour examination paper in mathematics in which the candidate is expected to answer only six or seven extended questions.

The assessment of such examinations has been conducted in a largely pragmatic way. Rightly or wrongly, many of those involved in administering the national assessment systems have regarded classical psychometrics as having little to say about the design, implementation and assessment of such complex, performance-based tasks.

***Dylan Wiliam** (PhD) is Senior Lecturer in Mathematics Education at the Centre for Educational Studies, King's College, University of London, Great Britain.*

Note: This article is based on a paper presented at the Nordic symposium Research on Assessment in Mathematics Education in Göteborg, November 5-7, 1993.

However, over the last thirty years there has been increasing concern that even such complex performance-based examinations only assess a sample of the mathematics felt to be important. Other forms of assessment such as 'portfolios', extended pieces of work, and oral and aural tests were developed. These developments were given added impetus when, in 1984, the Government announced that from 1988, all syllabuses for the school-leaving examinations should incorporate some school-based assessment.

The same pragmatic approach to assessment has been applied to the co-ordination of these new school-based assessments, with little attention being paid to the underlying theoretical issues. This is unfortunate because a substantial amount of very important work has been carried out by the five regional Examination Groups that administer national examinations which has not been published.

This article is an attempt to pull the two poles closer together – by bringing to the attention of psychometricians some of the innovative assessment practices undertaken in Great Britain over the last thirty years, and by extending some of the concepts of psychometrics so as to be more applicable to the kinds of authentic assessments that are used in public examining.

In section 2, I will characterises the kinds of tasks that have come to be associated with school-based assessment in England and Wales, and describe some of the schemes that have been proposed for their assessment. I shall argue that these schemes, by treating certain approaches to supposedly 'open' tasks as canonical have tended to stereotype mathematical activity in classrooms, and that this problem is inherent in all prescriptive assessment schemes.

Section 3 reviews the development of the concept of the validity of an assessment, emphasising the role of inferences made on the basis of the assessment results rather than the assessments themselves. Section 4 takes this idea further by focusing on the referents of the assessment – ie with what is the observed behaviour compared? The well-known norm- and criterion-referenced interpretations of assessment results, and the less well known ipsative interpretations are discussed, but it is argued that these are inadequate to describe some of the assessment practice that has emerged in recent years. A fourth kind of assessment and interpretation – construct-referenced assessment – is proposed and elaborated.

Finally, in section 5, some of the practical requirements of construct-referenced assessment are discussed briefly, and the concepts of Signal Detection Theory are proposed as being adequate for the evaluation of the dependability of such assessments.

2 Assessing authentic tasks in mathematics

One of the earliest uses of authentic tasks in a 'high-stakes' assessment setting was provided by a version of the school-leaving examination offered in 1966. In one of the three examination papers (Associated Examining Board, 1966), candidates had four hours to answer just one question from five:

1 Discuss the relevance of matrices to networks. Illustrate by suitable examples.

2 Discuss "Relations" with special references to their representations. Illustrate by suitable examples.

3 Discuss the applications of sets to linear programming.

4 [After a definition and an example of a simple continued fraction] Investigate simple continued fractions.

5 Investigate either: Quadrilaterals: classification by symmetry, or: Triangles and their associated circles.

Over the next twenty years, a variety of syllabuses incorporated such 'open-ended' tasks, either as part of a school-based examination component or as questions in a formal examination. However, the means for their assessment were generally ad hoc, and often highly idiosyncratic.

The announcement in June 1984 that 20% of the assessment in all school-leaving examinations in mathematics should be school-based triggered a sudden upsurge in the development of authentic tasks in mathematics, although the development has tended to favour some kinds of tasks more than others. In practice, there has been a strong bias towards open-ended problems in 'pure' mathematics, especially those involving combinatorics and enumeration.

Wells (1986) argues that as well as being stereotyped in terms of content, the approach to these tasks in classrooms is also stereotyped – an approach Wells calls data-pattern-generalisation (DPG) – relying on a 'naive inductivist' view of mathematics.

The tasks that have been developed in the last thirty years therefore cannot claim to be a comprehensive or representative selection of mathematical activity. Nevertheless, the open-ended nature of the tasks did create significant problems for the constructors of assessment-schemes. The use of the results of these examinations to make life-affecting decisions about employment and further education required that the assessments be highly reliable, while at the same time, the open-ended nature of the tasks makes it very difficult to prescribe and cater for the likely responses.

The schemes that have been developed are either generic where a general assessment scheme is used for more than one (and, typically, every) task or task-specific where a separate assessment scheme is constructed for each task. These are discussed in turn below.

2.1 Generic assessment schemes

Clearly, if workable generic schemes could be produced, they would be much more efficient than having to re-author new assessment schemes as new tasks were developed. Accordingly, almost all of the effort of the five regional Examining Groups responsible for the development and implementation of the new examination went into the development of generic assessment schemes. A large number of assessment schemes were produced, and a retrospective analysis of these schemes suggested that almost all schemes focused exclusively on one of two aspects of increasing competence in mathematical investigations (Wiliam, 1989).

In the 'cognitive demand' approach, the focus (adopting a metaphor from competitive diving) is on the 'degree of difficulty' of the task. Continuing the diving metaphor, the other approach focuses on the extent of progress made on the task or the 'marks for style'.

2.1.1 'Cognitive demand' approaches

The developmental psychology literature provides a number of models of the 'difficulty' of a task (Case, 1985; Pascual-Leone, 1970; Piaget, 1956), based on the cognitive processes required for successful action. Other models, such as the SOLO taxonomy (Biggs & Collis, 1982) ignore the cognitive processes and concentrate instead on the quality of the learning outcome.

However, it was clear from very early development work that these frameworks were unlikely to be useful in developing assessment schemes for open-ended work for two reasons. The first was that the number of different stages in the frameworks tended to be small – typically only four or five levels to cover the whole range of intellectual development from birth to adulthood – while the new examination was to report on an eight-point scale the attainment of just the age-16 cohort.

The second problem was that even though the number of levels was small, many tasks that should, according to their structure, be at the same level, actually showed widely differing degrees of difficulty – a phenomenon dubbed 'horizontal décalage' by Piaget & Inhelder (1941).

2.1.2 'Extent of progress' approaches

Instead, almost all the examination groups attempted to assess authentic tasks in terms of a problem-solving heuristic. Based on Polya's (1957)

four-steps of problem-solving, John Mason and Leone Burton had developed models of the problem-solving process (Burton, 1984; Mason, 1984) which were developed further by the Examining Groups into viable assessment schemes.

The schemes have used 'criteria' to exemplify the stages reached by students although these are not the precise behavioural objectives advocated by proponents of criterion-referenced assessment such as Popham (1980). Instead they used broad descriptors such as:

"Formulates general rules" (LEAG, 1987, p. 6),
"Express a generalisation in words" (OCEA, 1987, p. 13),
"Make and test generalisations and simple hypotheses" (DES & Welsh Office, 1989, p. 4)
"Make a generalisation and test it" (DES & Welsh Office, 1991, p. 4)

All these process-based schemes have, by ignoring the task variables, treated all tasks as essentially equivalent. Consequently, these process-based schemes would not distinguish between the same process displayed in different problem-contexts, even though the difficulty (as determined by, say, facility) might be very different.

For example, the number of integer-sided triangles that can be made with longest side n is given by

$$\frac{n(n+2)}{4} \text{ for even } n \text{, and } \frac{(n+1)^2}{4} \text{ for odd } n,$$

while the number of such triangles with perimeter n is

$$\frac{n^2+6n-1+6(-1)^{\frac{n+1}{2}}}{48} \text{ for odd } n \text{ not divisible by 3, and}$$

$$\frac{n^2+6n+15+6(-1)^{\frac{n-3}{6}}}{48} \text{ for odd } n \text{ divisible by 3,}$$

with the results for even n the same as for (odd) $n-3$ (Wiliam, 1993b). Yet both could arise naturally out of the open stimulus "Investigate integer-sided triangles".

The descriptor "Formulates general rules" mentioned above was associated with the highest grade of the new examination – the General Certificate of Secondary Education or GCSE – and was designed to be attained by about 5% of the age – 16 cohort. The generalisation for n as longest side would probably be regarded as 'too easy' for this standard, and that for n as perimeter as too hard. In this sense, the heuristic-based scheme does not give what teachers and examiners would regard as the 'right' result for either of these two versions of the same open task, even

where the response of the student follows the model of progression implicit in the assessment. Other approaches taken by students might diverge even further from the progression envisaged by the constructors of the assessment scheme.

Teachers' school-based assessments are subject to scrutiny by external moderators, who have (and have used) the power to revise marks by coarse re-scaling. Aware of this, and concerned to avoid having their school-based grades revised downwards, it appears that teachers have 'played safe', and used only coursework tasks that conform to the model of progression and the particular calibration implicit in the generic descriptors. This has produced a considerable stereotyping of the kinds of open-ended mathematical tasks that teachers offer to students – typically combinatoric problems with two independent variables and quadratic generalisations.

This convergence can be viewed as teachers taking control of the system and 'making it make sense'. But can also be viewed from a Foucauldian perspective as legislating the horizontal décalages out of existence by constraining the discourse within which the assessment takes place (Foucault, 1977) – 'if it doesn't fit the scheme it's not proper mathematics'.

These stereotyping effects would appear to be inherent in any generic assessment scheme. Nevertheless, such a model of general grade or level descriptors has been adopted for one of the dimensions of the revised National Curriculum for mathematics (DES & Welsh Office, 1991), nominally accounting for 20% of the curriculum.

2.2 Task-specific schemes

In response to the difficulties raised by horizontal décalage when using generic level descriptors, some assessment schemes (Bell, Burkhardt, & Swan, 1992; Graded Assessment in Mathematics, 1988, 1992) have developed task-specific level descriptions which take into account the context and difficulty of the tasks.

This creates an immediate difficulty in that only tasks for which assessment schemes have been prepared can be used, thus limiting the range of tasks than can be used. However, in the development of the Graded Assessment in Mathematics (GAIM) scheme at King's College during 1984-1990, we discovered another difficulty. When performance descriptions for particular tasks were presented, teachers often regard the identified descriptions as the only way of achieving a particular level, rather than an exemplification of the standard associated with that level. Furthermore, this was only partially alleviated by presenting multiple descriptions for each exemplified level. As well as restricting the kinds

of tasks that can be used, therefore, task specific schemes may well promote stereotyped approaches to tasks to a greater extent than is the case with generic assessment schemes.

Open tasks are claimed, by their proponents, to allow opportunities for students to pose problems, to refine areas for investigation, and to make decisions about how to proceed (Brown & Walter, 1983; Mason, Burton, & Stacey, 1982). However, while the activity may be student-centred, the assessment is not. From a Foucauldian reading of the situation, it is clear that the discourse is constrained. In a high-stakes setting (Popham, 1987, p. 77), or one in which the stakes are perceived to be high (Madaus, 1988, p. 86), students will be 'disciplined' (Paechter, 1992) into adopting easily assessable, stereotyped responses. They will still be playing 'Guess what's in teacher's head'.

To sum up, by delineating particular 'canonical' responses, the task-specific schemes appear to lead teachers to direct students towards approaches that yield more easily 'assessable' responses. On the other hand, general schemes have tended to treat all tasks as equivalent, with scoring dependent upon the mathematical processes involved, which has placed a premium on selecting tasks that are likely to elicit the appropriate processes.

What is required is a way of assessing authentic tasks on their own terms – in terms of what the student set out to do, but it does not seem as if any kind of explicit assessment scheme can achieve this. The possibilities for an implicit assessment scheme are discussed in sections 4 and 5, and in order to lay the foundations for this discussion, section 3 reviews the important developments in the concept of the validity of assessment.

3 The validity of assessments

The classical definition of validity has changed little over the last 50 years. Garrett (1937) defined the validity of an assessment as the extent to which it measured "what it purports to measure" (p. 324) although applying this in practice has led to a proliferation of kinds of validity.

Clearly it is important that an assessment is both relevant to the domain addressed and representative of it, and these two requirements are traditionally regarded as defining the extent to which the assessment has content validity. Unfortunately, the use of the term 'content' in this way appears to exclude assessment of the psychomotor and affective domains. Popham (1978) proposed but then retracted (1980) the term 'descriptive validity' as more appropriate and Embretson (1983) has used the term 'construct representation' to focus on the relationship between the psychological processes involved in responding to the assessment.

However, while assessments are frequently used to determine an individual's competence in a particular domain, they are more often administered in order to make decisions. Sometimes, the decision relates to future performance ("can this person become a pilot?") in which case the predictive aspect of validity is paramount. At other times, whether the assessment produces results similar to a more complex procedure is more important ("is this child being sexually abused?"), so that validation is concerned with the concurrence of the two procedures. Both predictive validity and concurrent validity concern the ability of an assessment to stand as a proxy for performance on some criterion, and are therefore often referred to collectively as criterion-related validity.

The idea that validity should be more concerned with the inferences made from assessment results, rather than the results themselves was made more explicit in the development, during the 1950s, of construct validity.

The term construct validation was first introduced in 1954 by a joint committee of the APA, AERA and NCME (1954) and required "investigating what psychological qualities a test measures, ie by demonstrating that certain explanatory constructs account to some degree for performance on the test" (p. 14). Originally intended to be used only where "the tester has no definitive criterion measure of the quality with which he is concerned and must use indirect measures to validate the theory" (p. 14), construct validation was held by Loevinger (1957) to be "the whole of validity from a scientific point of view" (p. 636). This view was disputed by Bechtoldt (1959) who regarded the definition as conflating the meaning of a score – construct representation – with what it signifies (nomothetic span). Nevertheless, by the late 1970s, Angoff (1988) notes that Loevinger's view "became more generally accepted" (p. 28), and Messick (1980) asserted that "construct validity is indeed the unifying concept of validity that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships" (p. 1015).

In the same paper, Messick argued that even this unifying concept of construct validity was not broad enough to allow a full consideration of the quality of an assessment. In addition to construct validity, he argued that the evaluation of an assessment procedure should consider:

a) the value judgements associated with the interpretations of the assessment results
b) evidence for the relevance of the construct and the utility of the particular applications, and
c) the social consequences of the proposed assessment and the use of the results – what Madaus (1988) has called the impact of an assessment.

This was presented as the result of crossing the basis of the assessment (whether the focus is on the evidence provided by the assessment, or on the consequences of its use) with the function of the assessment (whether the focus is on the interpretations of the assessments or their application). This provides the structure for Table 1. In addition Table 1 includes some of the other terms proposed by other writers which, while not being exact matches to Messick's partitioning of types of validity argument, appear to be close enough to be instructive.

	result interpretation	result use
evidential basis (Messick) interpretive basis (Moss)	construct validity (Messick) construct representation (Embretson) meaning (Bechtoldt)	construct validity + relevance/utility (Messick) nomothetic span (Embretson) significance (Bechtoldt)
consequential basis	value implications	social consequences (Messick) impact (Madaus)

Table 1. Facets of validity.

The various reasons given by different authors for incorporating authentic tasks into a scheme of assessment can be located within this framework. Incorporating authentic tasks into an assessment scheme for mathematics because it is believed to represent better the mathematical performance of the student is an appeal to construct representation (top left). Doing so because we believe we can better predict success in further study or employment is concerned with relevance or utility as well construct validity (top right). Doing so because it shows investigative work to be an important part of mathematics and so worth assessing concerns the consequential basis of result interpretations or, in other words, the value implications (bottom left). Finally, if we assess investigations because we believe that this will encourage teachers to incorporate such activities into their teaching, then we are concerned with validation in terms of the social consequences of the assessments (bottom right).

The important point about this framework is that while interpretive validity arguments (ie top row) can (although need not) be discussed

within a rationalist programme, arguments that involve discussion of consequences (bottom row) must be conducted within a value-system.

The incorporation of the impact or social consequences of assessments into validity argument has enriched the field enormously, and has illuminated many aspects of practice that could not easily be explained with narrower concepts of validity. In particular it can be used to illustrate the role that values play in assessment.

In one reading, teaching to the test is unacceptable (or even morally 'wrong') because it robs a test of its ability to provide useful information (Cannell, 1987). Other readings are however possible. In the US, given the use that is to be made of test results, many have argued that, at times, not teaching to the test is more damaging to students than doing so (Airasian, 1987).

As well as discussion of these negative 'backwash' effects, there has recently been a substantial amount of interest in the possible beneficial effects of the assessments, leading to the idea of the 'beneficence' (Elton, 1992) of an assessment. As an example of this, in 1988, the GAIM project described itself as an 'assessment-led curriculum development' project and was premissed on the assumption that teachers' practice could be changed by changing summative assessment practices. In the US, similar arguments are made by proponents of 'measurement-driven instruction' or MDI (Airasian, 1988).

It is clear, therefore, that the unified concept of validity is, ultimately, subjective and personal: "a test is valid to the extent that you are happy for a teacher to teach towards the test" (Wiliam, 1992, p. 17).

4 Norms, criteria and other referents

The previous section presented a framework within which the incorporation of authentic tasks in mathematics assessments can be considered. However, very few of the benefits of such tasks are likely to accrue if they are subject to the kind of stereotyping that, it was argued in section 2, is likely to occur if prescriptive schemes of assessment are used. This section suggests that a solution lies in a movement away from traditional notions of criterion- and norm-referenced assessment, and towards the idea of a 'construct-referenced' assessment.

Any assessment functions by comparing the behaviours observed during the assessment with something else – the referent. This referent might be the performance of other students of the same age, an external criterion, or even the student's own previous performance. The importance of the referent is that the interpretations of assessment results are (implicitly at least) inferences with respect to these referents. It is for this reason that it is now widely acknowledged that it is more useful to

speak of norm- and criterion-referenced interpretations of assessment results, than of norm- and criterion-referenced assessments per se. However, the design of the assessment still needs to take into account the inferences that it is proposed to make from the outcomes.

While it is possible, in some cases, to make satisfactory criterion-referenced interpretations from tests designed to provide norm-referenced inferences (and vice-versa), this is not always possible, and constructors of assessments need to be bear in mind the interpretations that are likely to be made of assessment results. For this reason, it does still make sense to talk of a 'norm-referenced test', although what we mean is a test designed specifically to provide norm-referenced inferences or interpretations.

4.1 Norm-referenced assessment

In a norm-referenced assessment, the referent is the normative group, but the inferences that are sought are likely to be to a much wider population. The key is therefore the ability of the normative group to represent the population.

Typically, the interpretations are expressed (even if only implicitly) in terms of the proportion of the cohort doing better or worse than the individual in question, and, typically, the nomothetic span of the assessment is at least as important as its construct representation, placing a premium on the ability of the assessment to discriminate. A useful touchstone is that an assessment scheme has some degree of norm-referencing if 'sabotaging' the efforts of other candidates is likely to help your own assessment!

4.2 Criterion-referenced assessment

The term 'criterion-referenced assessment' is generally attributed to Glaser (1963), although the underlying ideas are much older. Writing nearly twenty years earlier, Cattell (1944), had suggested that as well as 'populometric' or 'normative' measurement, there was another category of 'absolute' measurements, which related performance to "literal, logical dimensions defining events" (p. 295). He termed this assessment 'interactive' to stress that what was being related was the interaction between the individual and the "external world" (p. 294). The essence of criterion-referenced assessment is that the domain to which inferences are to be made is specified with great precision (Popham, 1980).

However, as Angoff (1974) has pointed out, if you scratch the surface of any criterion-referenced assessment, and you will find a norm-referenced set of assumptions lurking underneath. Popham (1993) regards

this not just a property of poorly-framed objectives, but an inevitable feature of all performance criteria. Any criterion will have a degree of 'plasticity' (Wiliam, 1993c, p. 342) in that there are a range of interpretations that can reasonably be made.

The particular interpretation of a criterion that is chosen should be the most useful bearing in mind the inferences that are desired. The central feature of a criterion-referenced assessment that distinguishes it from norm-referenced assessment is that once we have decided on an interpretation, it doesn't then change according to the proportion of the population achieving it.

4.3 Ipsative assessment

In the paper cited above, Cattell also proposed a third form of measurement "for designating scale units relative to other measurements on the person himself" (p. 294) which he termed ipsative measurements (from the Latin: ipse = self). Authors such as Stricker (1976) have further required that with ipsative measures, "the sum of scores for a set of variables is the same for each person" (p. 218).

More recently, at least in the UK, there has been a tendency to use the term ipsative assessment to describe the performance of an individual at one moment in time, compared with that individual's previous levels of performance (eg Stronach, 1989). Such comparisons are certainly not norm-referenced, since there is no comparison with a normative group, nor do such assessments satisfy the requirements of domain-specification needed in criterion-referenced assessment.

However, such assessments are part of the day-to-day activities of good teachers – arguably the most significant part – and the relative lack of attention that these kinds of assessments have received in the literature hardens many practising teachers' beliefs that psychometrics has nothing useful to say to them.

4.4 Construct-referenced assessment

There is another class of referents, widely used in education, to which assessments are frequently related, that have received little attention in the psychometric literature. These referents are used when the domain of assessment is holistic, rather than being defined in terms of precise objectives.

The essence of this fourth kind of referent is that the domain of assessment is never defined explicitly. Examples of behaviours are used in illustrating or exemplifying performance, but the standards, in that they exist anywhere, exist as shared constructs in the minds of those involved in making the assessments.

The most successful example of this kind of assessment in the UK has been the assessment of GCSE English, which, in three-quarters of the schools in England and Wales, is entirely school-based. In order to safeguard standards teachers are trained to use the appropriate standards for marking by the use of 'agreement trials'. There are many models for such agreement trials, but typically, a marker is given a piece of work to assess. 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 term domain-referenced assessment might be an appropriate description for this kind of assessment but for the fact that most authors – see for example Berk (1990, p. 490) or Hambleton and Rogers (1991, p. 4) – use this synonymously with 'criterion-referenced' assessment. Because of this, and because of the use that is made of shared constructs, I have proposed the term 'construct-referenced assessment' (Wiliam, 1992) – first used by Messick (1975) – to describe this kind of assessment.

The meaning I wish to attach to the term has a slightly different emphasis from that proposed by Messick, but both emphasise the meanings that are attached to assessment results (Bechtoldt, 1959) or how well they represent the constructs (Embretson (Whitely), 1983). The difference is that in Messick's terms, the constructs exist either in terms of traits within the individual or in terms of concordances with a nomological network within which the construct is defined, while the definition proposed here is essentially social. It is not necessary that the construct exists within a nomological network; merely that raters share the construct to a sufficient extent that they exhibit enough agreement about their ratings for the purpose in hand. The assessment is not objective, but the UK experience 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.

Both criterion-referenced and construct-referenced assessments relate performance to some external standard, and although there may be cases where both are equally appropriate, it will usually be the case that one is more appropriate than the other.

Where aspects of performance or achievement can be broken down into specific behaviours, which collectively exhaust the domain, then that performance is reducible. In such a case we are assured that if someone can perform each of the constituent behaviours, then we also know that they can perform the complete behaviour. Where the specific

behaviours can be defined precisely by explicit criteria, we can say that the behaviour is definable. Criterion-referenced assessments are appropriate for achievements that are both reducible and definable.

However, many complex skills cannot be treated in this way; the whole may be greater than the sum of the parts (so the performance is not reducible), or we cannot write down criteria which capture what we want to describe (so the performance is not definable). This is recognised in many areas of social decision making, and is encapsulated in the often-cited legal dictum that 'hard cases make bad law'.

Construct-referenced assessments may use statements to describe the domains – indeed many systems of assessment describe these statements as criteria – but the role of these statements is quite different from the role of criteria in criterion-referenced assessments.

The touchstone for distinguishing between criterion- and construct-referenced assessment is the relationship between the written descriptions 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.

In practice, no assessment relates exclusively to a single one of these four kinds of referents. For example a selection procedure for employment is likely to combine several aspects. If there is only a single post, then to be successful, one has to be the best candidate (in some sense) so there is a degree of norm-referencing. However, one also has to be 'appointable'; if there are selection criteria, then this may be a criterion-referenced assessment, but if there are not, then it may involve some construct-referencing (ie can this person do the job?). To complete the analogy, ipsative concerns may be involved if the application is for promotion, where the decision might be about whether the candidate has made sufficient improvement over (say) the last year.

5 Implementing construct-referenced assessment

Although the use of the term may not be well-established, there is a considerable amount of experience in England and Wales of the development of construct-referenced assessment. Unfortunately, this has not been researched rigorously nor has it appeared in the literature of educational and psychological measurement. The findings presented below must therefore be considered as suggestive at best, but in the absence of any more rigorous literature, may help others avoid some of the problems solved by trial and error in the UK.

5.1 Assessor training

The major requirement in achieving a dependable system of construct-referenced assessment is achieving unidimensionality: teachers may disagree about which grades, levels or scores to award, but they must agree on the rank order. For example, a common experience in the early development of school-based assessment in GCSE English was that some teachers focused unduly on the technical accuracy of the writing, almost ignoring the descriptive quality. Others paid little attention to punctuation, grammar and the 'conventions' of Standard English, awarding grades almost solely on the basis of descriptive power.

The first stage is to identify the various dimensions of competence in the domain, however intuitively, and, within each dimension, to identify (again possibly intuitively) degrees of progression. Samples of students' work for agreement trialling can then be selected to illustrate the extremes of differences in grade or quality along the different dimensions.

Once the illustrative set of work samples has been determined, there are many ways to proceed. For example, consistent rank ordering of samples can be established first, with correct gradings established as a subsequent calibration exercise, or both can be established simultaneously.

It is our experience that in many cases, teams of experienced teachers working together can produce unanimous agreements quickly, presumably because of the amount of experience they share, although how quickly new entrants to the profession can be enculturated is an important issue. However, preliminary results from work with mathematics students on initial teacher training courses (Gill, 1993) has shown that constructs of 'levelness' can be formed quite quickly. This is, however, a single, very small-scale study in a complex and diverse field and an exploration into how this process of enculturation can be speeded up must be a pressing item on the research agenda of construct-referenced assessment.

The Graded Assessment in Mathematics scheme mentioned above developed a series of mathematical investigations and practical problem solving tasks which had to be graded on a scale from 1 to 15, with the top seven grades being equivalent to the grades of the school leaving examination at 16 (the GCSE). Although very little of the project's findings have been published, several lessons about communicating constructs of assessment standards to teachers were learnt (Wiliam, 1993a):

- generalised level descriptions caused too many 'décalages' to provide a reliable basis for assessment.

- task-specific level descriptions tended to be either too specific to a particular approach taken to the task, or too general for teachers to be able to identify which were the important features. There did not appear to be an easily located 'middle ground'.
- attempts to communicate standards were consistently most successful when actual samples of students' work for each of the levels of the system, annotated to illustrate important points, with several different approaches to each task at each level, were used.

Generating such samples of work, however, caused its own problems. Whenever teachers were asked to provide specific samples of work to exemplify levels, the examples provided were usually at the correct level, but extremely well presented. To use such materials as exemplification would have created an unrealistic expectation of the standards associated with a particular level. Generally, more appropriate exemplification materials were generated when teachers were asked to submit whole class-sets of work, from which samples could be chosen.

5.2 Evaluating the quality of assessments

Construct-referenced assessment involves making complex, holistic and discrete attributions, and it is clear that in such settings, classical test theory is inappropriate for evaluating the quality of the assessments made. However, since any discrete attribution can be treated as a series of ordered dichotomous attributions, then signal detection theory provides appropriate indices of the accuracy of assessments.

Signal detection theory (SDT) developed out of attempts to analyse the performance of different (human) receivers of radio signals. However, while having been developed in communication engineering, the idea of dichotomous decision-making in the presence of noise has a very wide range of application (Green & Swets, 1966).

For consistency with the language of signal detection theory, the term 'positive' is generally used to denote the situation where the threshold has been exceeded, irrespective of whether this has positive or negative connotations (Swets, 1988, p. 1285). Similarly, the term 'negative' is used to denote a situation where the threshold is not reached. If the number of false and true attributions are expressed as a proportion of the true positives, then these proportions will sum to 1, as will the false and true attributions of true negatives. The behaviour of the system can then be described by two indices: the number of correctly attributed positives (called 'hits') and incorrectly attributed negatives ('false alarms') are usually chosen, although Sperling and Dosher (1986), argue that the use of hits and correct negatives gives more easily interpreted results.

This use of proportions satisfies Swets' first property required of a measure of the performance of a diagnostic system: that the measure should be unaffected by the proportion of positives and negatives in the test sample.

The second of Swets' requirements is that a measure of the performance of the system as a whole should be independent of the way that the decision criterion is set. In our case, we should want our measure of the accuracy of teachers assessing authentic tasks in mathematics to be the same whether they are told to be lenient or to be harsh in deciding whether to award a particular grade or level.

The essence of signal detection theory is that the decision-consistency of the system is measured over a wide range of possible settings of the threshold between lenient and strict interpretations of the criterion.

The graph of the pairs of false-positive proportions (false alarms) and true positive proportions is called the ROC (originally 'receiver operating characteristic', but now often 'relative operating characteristic') of the system, and describes the accuracy of the system over different settings of the criterion. This information can then be given to those who have to determine the setting of the threshold so the performance of the system at the chosen setting is known reasonably well in advance. If a single index, rather than a curve on a graph, is required, then Swets (1988, p. 1287) suggests that the area of the graph below the curve (denoted A) can be used as an index of system performance, which ranges from 0.50, when the ROC is a diagonal line (corresponding to the situation where no discrimination exists) to 1.00 (where there are no incorrect classifications).

The lack of a sampling distribution of the index A creates difficulties, but nevertheless signal detection theory appears to hold considerable promise where essentially continuous data has to be reported in a dichotomous way.

6 Summary

The incorporation of open-ended authentic tasks in formal assessments of mathematics achievement can be justified on many grounds. It has been argued that the inclusion of such tasks means that assessments represent better the nature of mathematical thinking, have greater utility for selecting students for advanced study, and represent more appropriately the values of mathematics.

However, the inclusion of such tasks in high-stakes assessment creates significant problems. Natural justice requires a high degree of inter-rater agreement, which has in the past meant adopting very tightly

controlled assessment schemes. While defensible for more narrowly focused tasks, such assessment schemes cannot anticipate all the different directions in which students might proceed. Task-specific assessment schemes cannot therefore be used without compromising the rationale for introducing authentic tasks in the first place.

While generic assessment schemes may, in the future, offer workable solutions, at the moment not enough is known about the nature of progression and competence in mathematics to provide a scheme that allows a piece of mathematical investigation to be assessed 'on its own terms'.

In section 2 it was argued that the generic schemes that have been developed to date have focused on particular aspects of performance (whether degree of difficulty or extent of progress). It was also argued that any generic scheme will, by its very nature, serve to canonise certain aspects of performance at the expense of others.

As a partial (and possibly temporary) solution, construct-referenced assessment has been proposed as a way of working towards sufficiently high inter-rater agreement, without compromising the rationale for open-ended work in mathematics.

Whether construct-referenced assessment is sufficiently different from norm- and criterion-referenced assessment to be useful remains to be seen, but even if the term does no more than focus attention on the inadequacies of the contrastive rhetoric of norm- and criterion-referencing as descriptions of complex human judgements, then it will have served its purpose.

References

Airasian, P. (1987). State mandated testing and educational reform: context and consequences. *American Journal of Education, **95**(3)*, 393-412.

Airasian, P. (1988). Measurement-driven instruction: a closer look. *Educational Measurement: Issues and Practice* (Winter), 6-11.

American Psychological Association, American Educational Research Association, & National Council on Measurement Used in Education (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin Supplement, **51**(2) part 2)*, 1-38.

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

Angoff, W. H. (1988). Validity: an evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum Associates.

Associated Examining Board (1966). *Mathematics syllabus C paper II*. London, UK: Associated Examining Board.

Bechtoldt, H. P. (1959). Construct validity: a critique. *American Psychologist, **14***, 619-629.

Bell, A. W., Burkhardt, H., & Swan, M. (1992). Assessment of extended tasks. In R. Lesh & S. J. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp. 145-176). Washington, DC: American Association for the Advancement of Science.

Berk, R. A. (1990). Criterion-referenced tests. In H. J. Walberg & G. D. Haertel (Eds.), *The international encyclopaedia of educational evaluation* (pp. 490-495). Oxford, UK: Pergamon.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: the SOLO taxonomy* (Structure of the Observed Learning Outcome). London, UK: Academic Press.

Brown, S. I., & Walter, M. I. (1983). *The art of problem posing*. Philadelphia, PA: Franklin Institute Press.

Burton, L. (1984). *Thinking things through*. Oxford, UK: Basil Blackwell.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: how all fifty states are above the national average*. Daniels, WV: Friends for Education.

Case, R. (1985). *Intellectual development: birth to adulthood*. New York, NY: Academic Press.

Cattell, R. B. (1944). Psychological measurement: normative, ipsative, interactive. *Psychological review, **51***, 292-303.

Department of Education and Science, & Welsh Office (1989). *Mathematics in the National Curriculum. London*: Her Majesty's Stationery Office.

Department of Education and Science, & Welsh Office (1991). *Mathematics in the National Curriculum*. London, UK: Her Majesty's Stationery Office.

Embretson (Whitely), S. E. (1983). Construct validity - construct representation versus nomothetic span. Psychological Bulletin, ***93(1)***, 179-197.

Foucault, M. (1977). *Discipline and punish* (Sheridan-Smith, A. M., Trans.). Harmondsworth, UK: Penguin.

Garrett, H. E. (1937). *Statistics in psychology and education*. New York, NY: Longmans, Green.

Gill, P. N. G. (1993). Using the construct of "levelness" in assessing open work in the National Curriculum. *British Journal of Curriculum and Assessment, **3(3)***, 17-18.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: some questions. American Psychologist, 18, 519-521.

Graded Assessment in Mathematics (1988). *Development pack*. London: Macmillan Education.

Graded Assessment in Mathematics (1992). *Complete pack*. Walton-on-Thames, UK: Thomas Nelson.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Hambleton, R. K., & Rogers, H. J. (1991). Advances in criterion-referenced measurement. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 3-43). Boston, MA: Kluwer Academic Publishers.

Loevinger, J. (1957). *Objective tests as instruments of psychological theory*. Psychological reports, 3(Monograph Supplement 9), 635-694.

London and East Anglian Group for GCSE Examinations (1987). *Mathematics: centre-based assessment*. London, UK: London and East Anglian Group for GCSE Examinations.

Madaus, G. (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.

Mason, J. (1984). *Mathematics: a psychological perspective*. Milton Keynes, UK: Open University Press.

Mason, J., Burton, L., & Stacey, K. (1982). *Thinking mathematically*. London, UK: Addison-Wesley.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist, **30***, 955-966.

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

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research, **62**(3)*, 229-258.

Oxford Certificate of Educational Achievement (1987). *Mathematics: putting it into practice.* Oxford, UK: Oxford International Assessment Services Limited.

Paechter, C. (1992, August). *Discipline as examination/examination as discipline: cross-subject coursework and the assessment-focused subject subculture*. Paper presented at British Educational Research Association Annual Conference held at Stirling University. London, UK: King's College Centre for Educational Studies.

Pascual-Leone, J. (1970). A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychologica, **32***, 301-345.

Piaget, J. (1956). Les stades du development mentale chez l'enfant et l'adolescent. In P. Osterreich, J. Piaget, R. de Saussure, J. M. Tanner, H. Wallon, & R. Zarro (Eds.), *Le probleme des stades en psychologie de l'enfants*. Paris, France: Presse Universitaire de France.

Piaget, J., & Inhelder, B. (1941). *Le développement des quantités chez l'enfant*. Neuchâtel, France: Delachaux et Niestlé.

Polya, G. (1957). *How to solve it.* Princeton, NJ: Princeton University Press.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

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. (1987). Can high-stakes tests be developed at the local level? *NASSP bulletin, **71**(496)*, 77-84.

Popham, W. J. (1993, April). *The instructional consequences of criterion-referenced clarity.* Paper presented at Symposium on Criterion-referenced measurement – a thirty year retrospective at the annual meeting of the American Educational Research Association held at Atlanta, GA. Los Angeles, LA: University of California.

Sperling, G., & Dosher, B. A. (1986). Strategies and optimization in human information processing. In K. Boff, J. Thomas, & L. Kaufmann (Eds.), *Handbook of perception and performance New York*, NY: Wiley.

Stricker, L. J. (1976). Ipsative measures. In S. B. Anderson, S. Ball, & R. T. Murphy (Eds.), *Encyclopedia of educational evaluation: concepts and techniques for evaluating education and training programs* (pp. 217-220). San Francisco, CA: Jossey-Bass.

Stronach, I. (1989). A critique of the 'new assessment': from currency to carnival. In H. Simons & J. Elliott (Eds.), *Rethinking appraisal and assessment Milton Keynes*, UK: Open University Press.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, **240**(4857)*, 1285-1293.

Wells, D. G. (1986). *Problem solving and investigations*. Westbury-on-Trym, UK: Rain Publications.

Wiliam (Williams), D. (1989). Assessment of open-ended work in the secondary school. In D. F. Robitaille (Ed.), *Evaluation and assessment in mathematics education* (pp. 135-140). Paris, France: UNESCO.

Wiliam, D. (1992). Some technical issues in assessment: a user's guide. *British Journal for Curriculum and Assessment, **2**(3)*, 11-20.

Wiliam, D. (1993a, April). *Assessing open-ended problem solving and investigative work in mathematics*. Paper presented at Second Australian Council for Educational Research Second National Conference on Assessment in the Mathematical Sciences held at Surfer's Paradise, Australia.

Wiliam, D. (1993b). Paradise postponed? *Mathematics Teaching* (144), 20-23.

Wiliam, D. (1993c). Validity, dependability and reliability in national curriculum assessment. *The Curriculum Journal, 4(4)*, 335-350.

Bedömning av autentiska uppgifter: alternativ till rättningsmallar

Sammanfattning

Artikeln beskriver de typer av 'autentiska' uppgifter – vanligen öppna, 'rent' undersökande uppgifter – som använts i nationell utvärdering i England och Wales under de senaste trettio åren. Rättningsmallar för bedömning har klassificerats som antingen problemspecifika eller sammanfattande. I den senare kategorin har mallarna i sin tur klassificerats efter uppgiftens 'svårighetsgrad' eller efter den 'dellösning' som eleven åstadkommit, beroende på vad som betonats mest i bedömningen. Vidare presenteras en beskrivning av begreppet validitet som fordrar att man beaktar värdet av implikationer och sociala konsekvenser av olika bedömningssätt. Det argumenteras för att både uppgiftsspecifika och sammanfattande scheman medför att elevernas sätt att angripa uppgifterna blir stereotypa. Ett alternativt paradigm till norm- och kriterierelaterade tolkningar av bedömningar, som kallas 'konstruktionsrelaterat', anses bättre för bedömning av autentiska uppgifter. I artikeln framförs förslag till implementation av ett sådant system. Utgående från en teori för att upptäcka "signaler" föreslås lämpliga mått för att utvärdera precisionen i sådana bedömningar.

Författare

Dylan Wiliam är universitetslektor i matematikämnets didaktik vid Centre for Educational Studies, Kings College, University of London, Great Britain.

Adress

Centre for Educational Studies, Cornwall House Annex, Waterloo Road, London SE1 8TX, Great Britain.
