Issues in Setting Standards :

Establishing Comparabilities

Edited by

Bill Boyle and Tom Christie

Falmer Press
(A member of the Taylor & Francis Group)
London • Washington, D.C.

7
Standard-setting Methods for Multiple Levels of Competence

Dylan Wiliam

Summary

Although the National Curriculum of England and Wales has undergone several revisions since its inception in 1988, the main features of its assessment have remained largely unchanged. Namely:

- all pupils are assessed at the ages of 7, 11, 14 and 16 (the end of each of the four 'key stages' of education);
- the assessments are to be criterion-referenced;
- the results are based on both teachers' judgments and the results of external tests;
- the assessments of pupils [apart from those at 16] are reported on a scale of age-independent levels (Secretary of State for Education and Science, 1988).

During the development of National Curriculum assessment in England, Wales and Northern Ireland, a variety of innovative scoring models for the external tests were piloted. Many of these were inconsistent with 'good practice' in aggregation (Cresswell, 1994) and so, from 1994 on, the level achieved on the test is to be determined entirely by unweighted summation of marks and the teachers' judgments and the test levels are to be reported separately.

This decision returned National Curriculum assessment to the 'mainstream' of educational and psychological measurement, but has also raised some new challenges. The problem of setting cut-scores for minimum competency tests is well-addressed in the literature, but there is very little about how multiple cut-scores can be set for different levels of competence, as required in National Curriculum assessment.

This chapter discusses some of the standard-setting methods that have been proposed for multiple levels of competence and describes and evaluates the results of some recent attempts at standard-setting for National Curriculum tests in mathematics and science.

Introduction

The term 'standard' is used in a variety of ways. Hambleton (1978) gives this example:

> School district A has set the following target. It desires to have 85 per cent or more of its students in the second grade achieve 90 per cent of the reading objectives at a standard of performance equal to or better than 80 per cent.

There are three kinds of standards involved in this example, relating to the proportion of:

1 items relating to a particular objective answered correctly by an individual student (80 per cent);
2 objectives to be achieved by each student (90 per cent);
3 students reaching the specified achievements (85 per cent).

Standard-setting in National Curriculum assessment involves aspects of all threeof these kinds of standards.

Allocating a student to one of the levels of the eight-level scale can be viewed as a series of dichotomous classifications—i.e. has the candidate reached level 4 or not? level 5? level 6? If achievement is measured simply by the total number of items answered correctly (so that any two candidates with the same total score are regarded as equivalent—essentially a 'Rasch' model) then standard-setting involves no more that setting a cut-score, and assigning mastery status to those that reach the cut-score, and non-mastery status to those who do not.

Although it may be possible to identify items in the National Curriculum tests with a particular level, since all correct answers are weighted equally, the levels serve primarily to stratify the domain from which the individual items are drawn. This means that the first and second kinds of standards are not distinguished in National Curriculum assessment.

The third kind of standard enters into the setting of the cut-score because of the design considerations involved in the ten-level scale. The original motivation for the ten-level scale required that, initially at least, most of the levels should be calibrated in terms of the median attainment of different age-cohorts (median 10-year-olds should be minimally competent at level 4, median 12-year-olds at level 5 and so on). While this will become less and less important as the standards 'float-free' of their original norm-referenced beginnings, paying attention to the third kind of standard ensures that distributions of scores are not too much at variance with public expectations.

The foregoing has discussed standards only in terms of particular fixed points. Wanting 85 per cent of students to achieve 90 per cent of a domain will place special emphasis around these particularly emphasized regions. It says nothing

about what proportion of the students would achieve 50 per cent of the objectives. Standards are usually discussed in unidimensional terms, even though the concept of standard involves a *distribution* of attainment throughout the population, and the shape of the distribution, apart from at the one point specified by the standard, is often ignored. So for example, the same *reported* 'standard' could mask quite different distributions of attainment, and this is particularly important when comparing populations from different countries or cultural groups. There is evidence, for example, that the 'tolerance of variability'—i.e. the extent to which a wide range of performance levels in a particular age cohort is acceptable—is much greater in the United Kingdom (UK) than, for example in the United States (US) and Japan (Robitaille and Garden, 1988; Schaub and Baker, 1991).

Standard-setting Procedures

In his 'consumer's guide' to standard-setting methods, Berk (1986) identifies thirty-eight different procedures, twenty-three of which are methods for setting standards and fifteen of which are procedures for adjusting them in the light of empirical data about their consequences. He classified the methods according to the degree to which empirical evidence about the effects of a particular choice of cut-score is fed into the procedure. In the past, methods have been described as either *judgmental* or *empirical,* but such a classification is unhelpful since all standard-setting procedures involve judgment, are essentially arbitrary (Hambleton, 1980), and cannot be discussed independently of a system of values (Messick, 1989). For this reason, in this paper, I will refer to *test-centred* methods, *examinee-centred* methods, and *combination* methods (Jaeger, 1989):

test-centred methods	no empirical evidence is used in determining the cut-score
combination methods	empirical data is used to complement or refine the choice of cut-score
examinee-centred methods	the choice of cut-score is based primarily on the distribution of scores, although judgmental factors feed into the definitions of, for example, 'minimum competency'.

Eleven of the twenty-three standard-setting methods described by Berk are 'test-centred', seven are 'combination methods', and five are 'examinee-centred'.

Cutting across this continuum from examinee-centred to test-centred methods is another which relates to the extent to which the standard-setting procedure is completed before the test is developed or implemented. For example, many assessments are *designed* to have cut-scores in the region of 50 per cent because a public expectation has been built up that suggests that this is where the 'pass-mark' will be. Other, 'mastery-oriented' assessments might be designed to have

cut-scores set at higher marks (say 70 per cent or 80 per cent), but the important idea is that the tests are designed with a particular cut-score in mind. For other tests, the primary requirement might be to ensure representative sampling from a domain, so that where the 'best' cut-score might be found is a secondary consideration. Put crudely, is the cut-score adjusted to fit the items, or are the items adjusted to fit the cut-score? Of course, these are not discrete approaches but opposite ends of a continuum; approaches involving pre-determining a cut-score might be described as *policy-oriented methods* and those that adjust the cut-scores in the light of the actual composition of the test, or the performance of candidates as *data-oriented methods*. These two dimensions (centering: examinee v test and orientation: policy v data) if treated dichotomously, give rise to four categories of standard-setting methods, which are discussed in turn below.

Policy-oriented Test-centred Procedures

In 1984, Sir Keith Joseph called for the results of school-leaving examinations in England and Wales to give far greater information about what a candidate could actually do (Secretary of State for Education and Science, 1984). Whatever his original intentions, the demand was interpreted as a requirement for the grades awarded in school-leaving examinations to provide detailed information about the skills and competencies it was reasonable to expect a candidate with a particular grade to have demonstrated in the examination. The difficulty was that the unreliability of all examinations necessitated some degree of compensation which then seriously weakened the inferences about particular skills that were warranted from the final overall grade[1]. The solution adopted throughout the UK was to 'stratify' achievement domains, so that certain parts of the domain were associated with particular grades or levels.

Such methods were investigated intensively during the 1980s during the development of 'grade-related criteria' for the national school-leaving examination (Good and Cresswell, 1988) and many graded or graduated assessment schemes (Foxman, Ruddock and Thorpe, 1989; Graded Assessment in Mathematics, 1992; Pennycuick and Murphy, 1988).

In some cases (e.g. mathematics) these strata were defined hierarchically so that succeeding strata subsumed earlier levels (psychological strata), while in others (e.g. science) the strata represented a partition of the domain (curricular strata). However, even where the domain has been partitioned (rather than nested), the curricular sequencing is based on an assumed notion of increasing difficulty, so that there is an imposed order among the components of the domain[2].

The curriculum model presented by such stratification of the domain is almost always reflected in the assessment model: each of the marks available in the test can be attributed unambiguously to one of the different levels. The fundamental technical problem in standard-setting, therefore, has been to develop a procedure for determining, for each candidate, which of the 'levels' or grades of the assessment scale best describes the overall performance of the candidate, given,

for each level, the proportion of items identified with that level that have been correctly answered by the candidate[3].

The most rigid 'hurdle' approach is to treat each level as independent, and to award the candidate the highest grade or level at which she has attained some fixed proportion of the available marks. Provided the proportion of marks required at each level is sufficiently high (75 per cent or 80 per cent is typical), users of test results are quite justified in concluding that a candidate awarded a particular level has demonstrated most of the achievements associated with that level.

Unfortunately, with such a procedure, a candidate who just fails to achieve the required proportion at each of the levels cannot be given a grade at all. To overcome this difficulty, Long (1985) proposed that marks should be added cumulatively, with multiple cut-scores being set on the single mark scale. Many such models have since been proposed, each of which represents, implicitly, a test-independent model of the performance of a 'minimally competent candidate' at each level. Long himself suggested that the cut score for level *n* should be set at 80 per cent of the marks at each of the tested levels up to and including level *n,* while Schagen and Hutchinson of the NFER proposed a cut-score of 100 per cent of all marks for tested levels below level *n* and 50 per cent of the marks at level *n*. Based on empirical data from National Curriculum science tests, Massey (1993) proposed that the cut-score for level *n* should be set by summing 10 per cent, 20 per cent, 30 per cent, 60 per cent, 70 per cent, 75 per cent and 80 per cent of the marks available at level *n+3, n+2, n+1, n, n−1 , n−2* and *n−3* respectively. These three models are illustrated in Figure 7.1.

The assumption that all the items associated with a particular level are equally difficult and equally discriminating appears, at first sight to be rather difficult to justify, but provided the models are known to test developers when they begin to develop tests, then fit to the assumed model can be one of the criteria used by developers in selecting items. However, the difficulty with such an approach is that the final test (and therefore the standard) is almost completely determined by the test developers. Although this is common in the UK, such a situation would be regarded as unacceptable in many countries, and as a result, many different test-dependent standard-setting procedures have been devised.

Data-oriented Test-centred Procedures

Several of the standard-setting methods discussed by Berk (1986) can only be used with multiple-choice methods. The only method that is applicable within the constraints of National Curriculum assessment (and which, fortunately, also gets the highest 'approval rating' in Berk's 'consumer's guide') is the procedure that has come to be known as the Angoff procedure (Angoff, 1971).

As usually applied, the Angoff procedure requires a team of judges, working independently, to assign to each item in the test a number between 0 and 1, corresponding to the probability that a 'minimally competent' candidate would

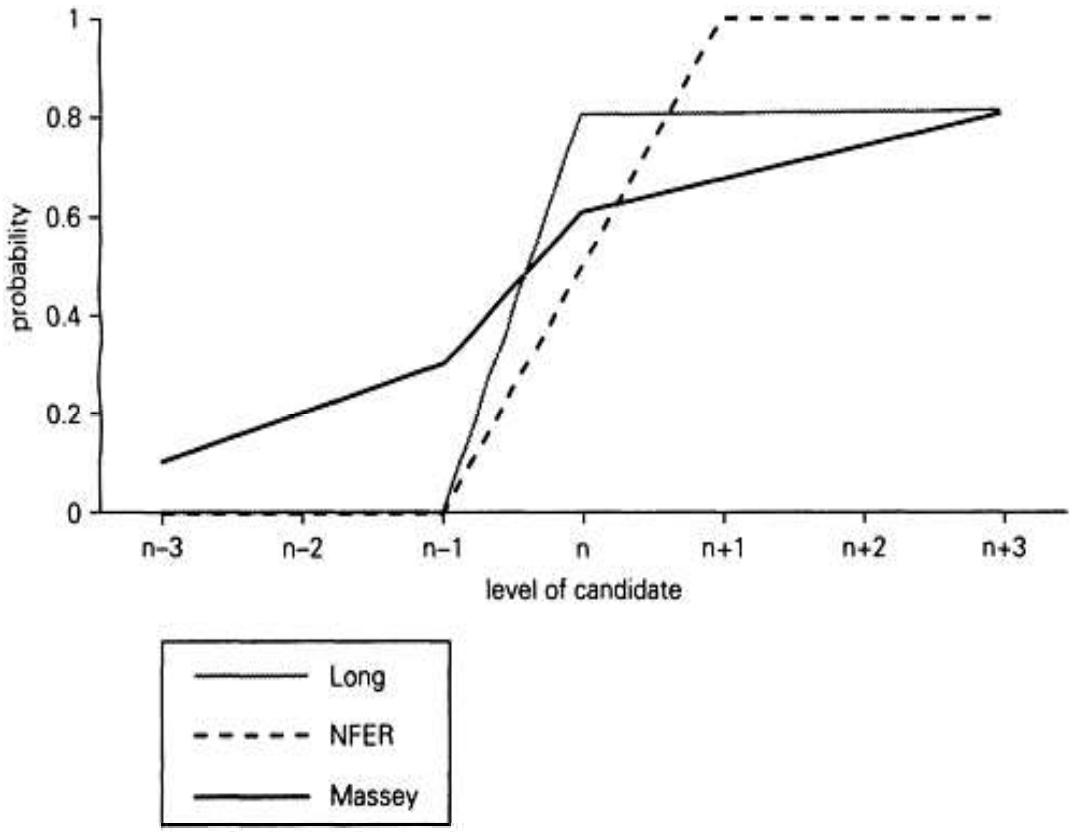

Figure 7.1: Probability of 'minimally competent' students' correct answers

be able to answer the question correctly. The judge's probabilities for each item are summed to give a minimally acceptable standard for that judge. The cut-score is then the average of the sums over all the judges.

The danger with such test-centred models is that they can generate standards which appear to be quite reasonable, but, in fact, are almost impossible to achieve. For this reason, there are many procedures that either use normative data explicitly in the original standard-setting, or allow empirical data to influence the cut-score.

Data-oriented Examinee-centred Procedures

According to Jaeger (1989), the most popular examinee-centred models in the US have been the borderline-groups and the contrasting-groups procedures proposed by Zieky and Livingstone (1977). The crucial point about such methods is that they require the allocation of candidates to groups (competent, borderline, inadequate) before the test is administered, and therefore call for some other method of assessment. The borderline-groups procedure defines a standard as the median test-score achieved by the borderline group, while the contrasting groups procedure sets a cut-score so as to maximize the

discrimination between the clearly competent and the clearly inadequate groups. The borderline-groups procedure therefore ignores the performance of all those judged to be either competent or inadequate, while the contrasting-groups procedure ignores all those deemed 'borderline'.

In England, models from decision theory have been used to inform the process of setting standards for the school-leaving university entrance examinations. Notable in this respect is the work of the Decision-Analytic Aids to Examining (DAATE) project (French, Wilmott and Slater, 1990).

Comparison of Methods

Many studies have been conducted in which different standard-setting methods are applied to the same test in order to establish the comparability of standards set by different methods. Although the data is difficult to summarize, it does seem that the contrasting-groups and Nedelsky (1954) procedures appear to generate more 'lenient' standards, and that the standard derived from the Angoff method may be the most demanding more often than would be expected by chance variation Jaeger (1989).

One finding does, however, emerge unequivocally: different methods produce different standards! Mills (1983) found that one method generated a standard that rejected nearly thirty times as many students as the other, and Jaeger's (1989) analysis showed that, averaged over thirty-two comparative studies, the cut-score suggested by the most demanding method was over five times the cut-score suggested by the least demanding method.

Glass (1978) interpreted the range of standards suggested by different methods as invalidating the whole idea of 'minimal competence'. Others (e.g. Hambleton, 1980) have argued that because different methods place emphasis on different aspects of performance, the discrepancies arise naturally out of the manifold interpretations that are placed on test results. A summary of the methods discussed is shown in Figure 7.2. No standard-setting method is ideal, but one may support the most important inferences that are likely to be drawn from the results better than the others. This suggests that ideally, several methods should be used, and the resulting cut-scores can be balanced with other (possibly not even measurement-related) factors, when determining a cut-score.

Applications to National Curriculum Assessment

In order to maximize reliability, some National Curriculum tests will be available in several 'tiers', with each tier covering three levels. Students estimated by their teachers to be at (say) level 6 would sit a test covering levels 5–7 so that the student could be awarded the level above or below that estimated by the teacher with a reasonable degree of reliability.

In National Curriculum assessment, an item relating to (say) level 5 on the eight-point reporting scale could, for example, appear in a test 'tier' covering

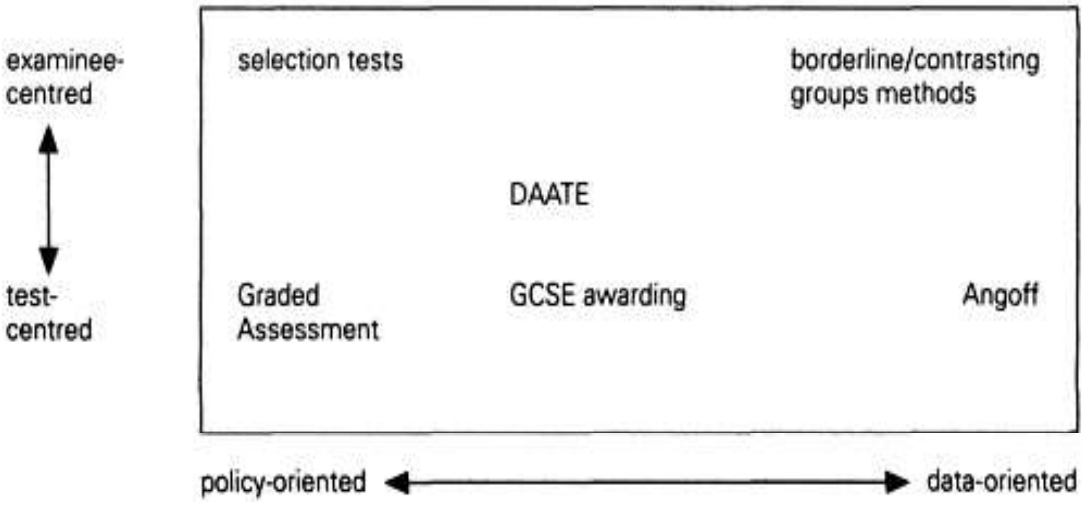

Figure 7.2: Classification of standard-setting procedures

levels 3–5, a tier covering levels 4–6 and in one for levels 5–7. With such a model of 'overlap' there would be five potential thresholds (in this case the 'minimally acceptable' standards for levels 3, 4, 5, 6 and 7)[4]. Each level 5 item would need to be rated with respect to each of these five thresholds, but once this is done, the same data is used for this item for each tier in which the item appears. In this way the item 'carries its difficulty rating around with it', thus avoiding complex technical problems of equating the standards on the different tiers.

In the original Angoff procedure, there is only a single threshold. Moving to multiple thresholds as required in National Curriculum assessment does not present any new theoretical difficulties, but there is one major practical difficulty.

With a typical National Curriculum test, the number of items and the number of thresholds against which each item is to be rated means that it is unlikely that all judges could rate all items against all thresholds. The question is, therefore, how should the items be allocated to judges? For example, if the judges involved in rating items against the level 4 threshold are different from those rating items against the threshold for level 5, there is the possibility that level 5 might end up being easier than level 4.

However, asking judges to rate the same item at more than one threshold also presents difficulties, since the Angoff method relies on judges holding in their minds an image of the 'minimally competent' student at the threshold in question. Whether judges can do this sufficiently well at more than one threshold, or change from one image to another quickly, is not clear. Several studies (two of which are summarized below) have already been conducted into the appropriateness of Angoff-type procedures for National Curriculum assessment, but much more research on the relative strengths and weaknesses of different experimental designs is needed.

The mathematics tests for 14-year-olds in 1994 covered levels 3–10 in five tiers: 3–5, 4–6, 5–7, 6–8 and 9–10. Early in 1994, twenty-five teachers participated in a two-day conference, using a three-stage Angoff procedure (i.e. initial judgment, followed by normative data, followed by discussion).

When the data were analyzed, a remarkably coherent pattern emerged. The cut-scores set tended to increase linearly according to level (about 20 per cent per level) and to decrease with the tier (so that the cut score for a particular level on one tier was, on average 23 per cent lower than the cut-score for the same level on the tier below. No significant differences were found between the two groups looking at the same levels, but the final cut-scores set were, on average, 6 per cent lower than those set initially (Lundy and Close, 1993).

More recently, Morrison and his colleagues at Queen's University of Belfast found that the normative data has very little influence on the standards set with the Angoff procedure (Morrison, Healy and Wylie, 1994). It seems, therefore, that the opportunity to discuss standards with colleagues is the most important factor in judges' tendencies to revise their initial standards downwards. This suggests that either a two-stage Angoff procedure (initial rating followed by discussion and revision), or a three-stage Angoff procedure, with normative data provided from pre-test and item piloting studies (rather than 'live' normative data) provide a sound basis for setting standards for National Curriculum assessment. Furthermore, in the Belfast study, the inter-rater effects accounted for only 11 per cent of the total variability in scores, confirming the findings from research relating to Angoff-based standard-setting in the US that panels of fifteen to twenty are quite adequate.

However, it is important to note that the foregoing discussion relates to standard-setting for individuals, rather than groups. Although the error attributable to the judges in the standard set is small in comparison to the test error, this will not be true for group data. Over a whole national cohort, the random error due to the test will be negligible[5], but the error due to the judges will remain exactly the same. So if in one year, the cut-score set is towards the lower end of the confidence interval, while the following year it is towards the upper end, then 'standards' within the population could appear to be falling simply due to chance variation in the cut score. This suggests that if year-on-year comparability is required, standard-setting panels should be as large as can be managed. However, the composition of such panels also needs careful consideration if the cut-scores are to reflect the progressive nature of the levels in the reporting scale.

Recommendations for Composition of Panels of Experts

Teachers will obviously be more expert at judging the standards with which they have more experience, but it is necessary that cut-scores relating to levels that are awarded at more than one key stage are set consistently. One way to secure this is to ensure that the expertise of panels setting standards for particular levels is in

	Level							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
KS1 teachers	27	26	12					
KS2 teachers	3	4	14	20	15	8	3	
KS3 teachers			4	10	15	22	27	30

Figure 7.3: Suggested constitution of panels of thirty for each level

proportion to the students who are awarded that level (this would, in effect, require that different teachers judge different levels, thus resolving one of the issues raised above).

For example, data on the distribution of attainment derived in Wiliam (1992) would suggest that panels of approximately thirty judges for each level should be constituted as shown in Figure 7.3. In practice, having just two panels—one composed of twenty Key Stage 1, eight Key Stage 2 and two Key Stage 3 teachers setting standards for levels 1, 2 and 3, and another composed of ten Key Stage 2 and twenty Key Stage 3 teachers setting standards for levels 4, 5, 6, 7 and 8 — would be a reasonable compromise.

Recommendations for Setting Cut-scores

National Curriculum assessment has a number of priorities all of which need to be given some attention in the setting of standards. However, rather than viewing these (often conflicting) priorities as requiring weights to be attached to them, it may be more helpful to view the priorities as concerns which need to be met. One way to think of these is as a needle on a dial: as long as the needle is not in the 'danger' or 'red' zone, then the standard is acceptable. Examples of such 'constraints' are 'If the cut-score for level 6 is below 57, you'll be able to get level 6 without getting any items correct at levels 6 or above' or 'If the cut-score's over 73, then you'll have to get some level 7 items as well as all the items up to level 6 right to get level 6'.

In order to take account of these multiple objectives for National Curriculum assessment, it seems appropriate that a series of aims be formulated that can be translated into tight criteria. These criteria can then be used to validate the chosen standard. Each criterion will yield a range within which the cut-score should fall. The cut-score would then be the lowest mark that simultaneously satisfies all the conditions. The actual aims and criteria to be used should be determined after consultation, but a preliminary list of four such is given below.

Aim 1: the standard should be based on a notion of minimum competence.
Criterion: The standard set should be at or below (but no more than two standard errors of measurement below) the cut-score determined by a two or three-stage Angoff procedure.

Aim 2: the standard for a particular level should relate only to items up to that level.

Criterion: The standard set for a particular level should not require any achievement at a higher level, nor be achievable without some achievement at that level.

Aim 3: the standard for a particular level should relate to the professional and public expectations for that level.

Criterion: The standard should result in facilities (assessed on the basis of the pre-test data) broadly in line with the facil ities suggested in the TGAT (1988) report. For example, the cut scores set should result in facilities for 14-year-olds in the following ranges:

93	level 3	100
86	level 4	96
69	level 5	79
45	level 6	55
21	level 7	31
4	level 8	14

Aim 4: the standard for a particular level in each subject should result in comparable facilities for all subjects.

Criterion: This will be achieved if a criterion similar to that pro posed for aim 3 is implemented. Otherwise some sort of similar criterion will need to be employed. However, it may already be too late. There is already a suggestion that the 'English creep' that resulted in the standards for English in school-leaving examinations being 'easier' than those for mathematics is now affecting National Curric ulum assessment!

Conclusion

There are other possible aims, and many might disagree with the operationalizations of those aims in terms of criteria presented above. Nevertheless, it seems to me that the principle of converting aims into criteria that yield precise intervals within which the eventually adopted cut-score must fall is useful. There is clearly no 'best' method, but by attending to a range of concerns, it may be possible to develop standards that pay reasonable heed to the conflicting demands made of National Curriculum assessment.

Notes

1 The use of profiles would of course have helped here but it was clear that there would still be a requirement to represent the overall performance in terms of a single 'brute grade' which would support inferences about particular aspects of performance.
2 While it would therefore be possible in a partitioned (rather than a nested) domain, for a student to acquire 'high-level' skills without necessarily acquiring 'low-level' skills, it is assumed that this will not happen due to the choice of curriculum sequences. Whether this assumption is warranted has been a source of much debate.
3 It is also possible, of course, that, in the interests of efficient test administration a candidate may not be assessed with respect to each level.
4 In general, with a 'regular' tiering strategy, with 1 level per tier, and p levels of overlap, the number of thresholds against which each item will need to be rated (t) is:
$$l + [(l - 1) / (l—p)] \cdot (l—p)$$
5 For example, in a cohort of 40,000 students, the standard error for cohort's mean would be half of one per cent of the error in each individual's score.

References

ANGOFF, W.H. (1971) 'Scales, norms and equivalent scores', in THORNDIKE, R.L. (Ed) *Educational Measurement,* Washington, DC, American Council on Education, pp. 508–600.

BERK, R.A. (1986) 'A consumer's guide to setting performance standards on criterion-referenced tests', *Review of Educational Research,* 56, 1, pp. 137–172.

CRESSWELL, M.J. (1994) 'Aggregation and awarding methods for National Curriculum assessments in England and Wales: A comparison of approaches proposed for key stages 3 and 4', *Assessment in Education: Principles, Policy and Practice,* 1,1, pp. 45–61.

DES/WO (1988) *Report of the Task Group an Assessment and Testing,* London, DES.

FOXMAN, D.D., RUDDOCK, G.J. and THORPE, J. (1989) *Graduated Tests in Mathematics,* Windsor, UK, NFER-Nelson.

FRENCH, S., WILMOTT, A.S.and SLATER, J.B. (1990) *Decision Analytical Aids to Examining: The DAATE Report,* London, UK, School Examinations and Assessment Council.

GLASS, G.V. (1978) Standards and criteria', *Journal of Educational Measurement,* 15, 4, pp. 237–261.

GOOD, F.J. and CRESSWELL, M.J. (1988) *Grading the GCSE,* London, UK, Secondary Examinations Council.

GRADED ASSESSMENT IN MATHEMATICS (1992) *Complete Pack,* Walton-on-Thames, UK, Thomas Nelson.

HAMBLETON, R.K. (1978) 'On the use of cut-scores with criterion-referenced tests in instructional settings', *Journal of Educational Measurement,* 15, 4, pp. 277–290.

HAMBLETON, R.K. (1980) 'Test score validity and standard setting methods', in BERK R.A. (Ed) *Criterion-referenced Measurement: The State of the Art,* Baltimore, MD, Johns Hopkins University Press, pp. 80–123.

JAEGER, R.M. (1989) 'Certification of student competence', in LINN, R.L. (Ed) *Educational Measurement,* Washington, DC, American Council on Education/ Macmillan, pp. 485–514.

LONG, H.A. (1985, 27 June-2 July) 'Experience of the Scottish Examinations Boards in developing a grade-related criteria system of awards', Paper presented at 11th Annual Conference of the IAEA held at Oxford, UK. Dalkeith, UK, Scottish Examinations Board.

LUNDY, I. and CLOSE, G.S. (1993) 'Experiences in setting cut-offs for key stage mathematics in 1994', unpublished report prepared for School Curriculum and Assessment Authority, London, UK, King's College London Centre for Educational Studies.

MASSEY, A. (1993) 'ZCriterion-related test development and national assessment standards', unpublished report prepared for School Curriculum and Assessment Authority, Cambridge, UK, University of Cambridge Local Examinations Syndicate.

MESSICK, S. (1989) 'Validity', in LINN, R.L. (Ed) *Educational Measurement,* Washington, DC, American Council on Education/Macmillan, pp. 13–103.

MILLS, C.N. (1983) 'A comparison of three methods of establishing cut-off scores on criterion-referenced tests, *Journal of Educational Measurement,* 20, pp. 283–292.

MORRISON, H.G., HEALY, J. and WYLIE, C. (1994) 'An investigation of the 1994 key stage 3 sample mathematics paper with Angoff-determined cut-scores at levels 3, 4 and 5', unpublished report prepared for School Curriculum and Assessment Authority, Belfast, UK, Queen's University of Belfast.

NEDELSKY, L. (1954) 'Absolute grading standards for objective tests', *Educational and Psychological Measurement,* 14, 1, pp. 3–19.

PENNYCUICK, D. and MURPHY, R. (1988) *The Impact of Graded Tests,* London, Falmer Press.

ROBITAILLE, D.F. and GARDEN, R.A. (Eds) (1988) *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics,* Oxford, UK, Pergamon.

SCHAUB, M. and BAKER, D.P. (1991) 'Solving the math problem: Exploring mathematics achievement in Japanese and American middle grades', *American Journal of Education, 99,* 4, pp. 623–642.

SECRETARY OF STATE FOR EDUCATION AND SCIENCE (1984) 'Speech to the North of England Conference, in Secondary Examinations Council Annual report 1983–84', London, UK, Secondary Examinations Council, pp. 60–68.

SECRETARY OF STATE FOR EDUCATION AND SCIENCE (1988) Assessment and Testing: A Reply to Mr Key, in Parliamentary Written Answers, Hansard, 7 June 1988, London, UK, Her Majesty's Stationery Office.

WILIAM, D. (1992) 'Special needs and the distribution of attainment in the National Curriculum', *British Journal of Educational Psychology,* 62, pp. 397–403.

ZIEKY, M.J. and LIVINGSTONE, S.A. (1977) *Manual for Setting Standards on the Basic Skills Assessment Tests,* Princeton, NJ, Educational Testing Service.

8
Seeing the Wood through the Trees: Setting Standards in the National Curriculum

Colin Robinson

Summary

The National Curriculum (of England and Wales) is primarily a medium through which the government attempts to involve itself in the education of future generations. For the first time in our history the requirements of what teachers are to teach and what children are to learn are being set out for all to see. The fact that they are so explicit, so public and, unfortunately, so open to different interpretations, makes the setting of standards a very important function. How do we go about it, and are we likely to succeed?

Introduction

In the beginning was the word. Not, in this case, the word of God, but the word of the curriculum Orders. Published over a period of some eighteen months, the Orders represented the combined wisdom of panels of experts: one in each of the ten subjects laid down by law as the Foundation subjects. The problem with words is that they are open to a variety of interpretations and they depend for their meaning on connections they make within the mind of the reader. This results in the same requirements being interpreted in different ways according to the assumptions and prejudgments made by the person interpreting them.

Take a simple requirement. Suppose we require that the pupil is taught to multiply two-digit numbers. What does this mean?

Take any two-digit numbers and you can develop strategies for multiplying them but with some there are different strategies that can be adopted: multiplying by 10, 20 etc. is very different from multiplying by 47. But is 10 x 47 any more difficult than 47 x 10? Does the arrangement of the numbers as follows:

10	or	47
× 47		× 10

make any difference? What is the effect of putting these numbers into a contextsuch as ‘How many people are there in forty-seven rows with ten people ineach?’ Does writing the numbers as words make a difference?

As teachers we know that all these differences do make a difference— at least in the early stages of learning. This does not mean that the children, when learning how to multiply, will always get the sums wrong. What is more likely is that they will be erratic—sometimes getting ‘difficult’ forms right whilst failing on what we thought were easy.

Progression

In setting out the original National Curriculum, too much credence was given to the idea that we can arrange learning in a hierarchy: starting with the easy things at level 1 and progressing through levels of increasing difficulty. Progression is the stuff of education. The whole of the learning process depends on the learners adding to what they know already. But learning does not follow the same pattern for every learner and, therefore, no single progression will adequately reflect the development of each individual.

Look at some of the hierarchical requirements for mathematics in the 1989 version of the mathematics curriculum—remembering that these statements of attainment defined the attainment target at each level and, as such, had the force of law.

Level 3 Pupil should

- *know and use addition and subtraction number facts to 20 (including zero);*
- *solve problems involving multiplication or division of whole numbers or money, using a calculator where necessary;*
- *know and use multiplication facts up to 5 x 5, and all those in the 2, 5 and 10 multiplication tables.*

level 4 Pupils should

- *know multiplication facts up to 10 x 10 and use them in multiplication and division problems;*
- *(using whole numbers) add or subtract mentally two 2-digit numbers; add mentally several single-digit numbers; without a calculator add and subtract two 3-digit numbers, multiply a 2-digit number by a single-digit number and divide a 2-digit number by a single-digit number;*
- *solve addition or subtraction problems using numbers with no more than two decimal places; solve multiplication or division problems starting with whole numbers.*

The point about all of these is that they are fine objectives for teaching. They do represent a sort of logical sequence that would make sense—you don’t try to

teach too much about the 7 or 9 times tables if the children haven't yet grasped the 2, 5 or 10. But if a particular child does grasp the 'more difficult' one first, we are just thankful for the bonus.

Assessment

The problem lies in trying to use these as criteria for assessment. Even if they were perfectly formed criteria which were wholly unambiguous in their meaning, they would still represent inappropriate assessment criteria. The linear progression that they assume is untenable—even within particular strands. If you start to require all of them before a particular level can be awarded, you produce wholly arbitrary results which can only be interpreted as representing the minimum standard attained—a very demotivating feature for the child who has achieved bits and pieces at many levels. So how have we tackled the problem?

Firstly we moved to an assessment regime that tests a very small sample of the domain—completely abandoning the notion of replicating classroom activities that the Task Group on Assessment and Testing put forward in its report (DES/WO, 1988). The sampling is too light to provide the necessary evidence of mastery of a particular statement, so the focus has shifted to the level as a whole.

But even in this shift, the level was still defined by the statements of attainment. The criterion for success was modified in the light of the number of statements tested. This led to even more arbitrary results. If the child happened to be successful on the statements assessed, the level was awarded. But there was no evidence that the statements within a level were sufficiently congruent to justify the assumption that success on those assessed could be taken as an indication that the child would have succeeded on the others. Perhaps even more worryingly, the fact that a child did not succeed on one or more of the assessed statements resulted in the level being denied, when assessment on the other statements might have led to success. Such a procedure could not be defended.

We therefore moved to the use of marks, albeit still trying to make use of the statements of attainment (which, after all, were the legal requirements). The statements were used as a basis for designing the questions and for establishing the cut-scores for the different levels. But now, if the pupil had achieved success on what was deemed a harder question, this could be set against the failure on an easier question. We were progressing towards a method familiar in most examinations.

The obvious next stage was to get rid of the statements of attainment. But they did have a value—as we have seen before—as objectives for teaching. They did set out the 'rules of the game' in a way that we had not previously attempted.

Revised Curriculum

When Sir Ron Dearing undertook his review of the National Curriculum last year, the whole question of the progressive ten-level scale came under scrutiny.

Opinion was divided but, in the end, the advantages for teaching were accepted and the scale was retained. What was agreed, however, was that the use of the statements of attainment should be abandoned. They have been replaced by what are called 'level descriptions' whose focus is the whole level, and which are to be used in a completely different way.

Statements of attainment set out to remove from the teacher (acting in the role of assessor) as much of the responsibility for making a judgment as possible. A child was assessed against criteria for the achievement of each statement and, if sufficient statements were achieved, the level would be automatically awarded. If the requisite number had not been 'ticked off, the level could not be awarded. Level descriptions rely on teachers making a judgment. They are to take the child's work as a whole and judge it in comparison with the level descriptions. There are likely to be features in the work that would be better described by the description at one level but others that make one believe that it is best typified by another. This change of use is signalled in the format of the new level descriptions. For example in mathematics at level 3:

> *Pupils have extended their understanding of place value to numbers up to 1000 and approximate these to the nearest 10 or 100. They have begun to use decimals and negative numbers, in contexts such as money, temperature and calculator displays. Pupils use mental recall of addition and subtraction facts up to 20 in solving problems. Pupils use mental recall of the 2, 5 and 10 multiplication tables, and others up to 5 x 5, in solving whole number problems involving multiplication and division, including those which give rise to remainders. They are beginning to develop other mental strategies of their own and use them to find methods for adding and subtracting two-digit numbers. They use calculator methods where appropriate. Pupils also solve problems involving multiplication or division of money by whole numbers.*

No longer do we have the 'bullet points' defining separate requirements. The continuous prose attempts to describe 'typical' performance, without requiring all of it to be displayed.

Setting Standards

So how do we go about setting standards? There is no doubt in my mind that the definition of standards in terms of a progressive scale across the whole of a child's school career is a step forward in recognizing what each part of the educational process is dependent upon and builds on what has gone before. It is therefore appropriate that we articulate that progression in some way, and the ten level scale is not a bad start (even if levels 9 and 10 are a bit difficult to envisage).

The next stage in the process is ensuring that a consensus emerges as to what constitutes performance at each level—balancing superior skills in one area against deficiencies in another to arrive at a level that really does do justice to the child's achievement. Two ways are suggested of achieving this: firstly, the circulation of examples of pupils' work assessed against the parameters of the new level descriptions with a rationale why the assessors decided to award each example a particular level. Secondly, and more importantly, standardization will come through practice: as teachers use their skills to assess their pupils, and exchange views as to the appropriateness of the assessments of their colleagues, a consensus will emerge. It may take a bit of time, but eventually there will be sufficient agreement as to what constitutes 'threeness' or 'fourness' for these to become the lingua franca of the staffroom. We must hope, however, that these terms will be attached to the attainments of the pupils and will not become labels for the pupils themselves, setting ceilings on expectations that have the opposite effect to the increase in standards that we all wish to achieve.

Reference

DES/WO (1988) *Report of the Task Group on Assessment and Testing,* London, DES.