National Foundation
for Educational Research

Maintaining Standards Over Time in National Curriculum English and Science Tests at Key Stage 2

A report for the Qualifications and Curriculum Authority

Paul Newton

March 2000

Contents

Executive Summary

The present report focuses upon procedures for maintaining national curriculum test standards over time. In doing so, it also addresses the issue of monitoring educational standards over time, as this is one of the three major uses to which national curriculum test results are currently put. Two other principal uses of test results are identified: monitoring the progress of individual pupils; and monitoring the progress of individual schools. The central argument of the report is that decisions concerning procedures for maintaining test standards must ultimately be guided by the intended uses of the test results.

The report begins with a brief discussion of the history of national curriculum assessment which shows how and why the original conceptual framework was flawed. The tests soon changed from being strongly 'criterion referenced' to being weakly 'criterion related'. This reflected an implicit acceptance of the inability of criterion referencing to provide an acceptable basis for standard setting in national curriculum tests. To focus the argument of the present paper, an alternative framework for understanding the maintenance of test standards is fleshed out. It is proposed that standards have been maintained when candidates with the same 'test potential' in different tests are rewarded with the same levels. Effectively, this amounts to the proposal that all factors that affect the performance of an individual from one test to the next should be recognised in marks (and, hence, levels) awarded, *with the sole exception of factors internal to the test itself.* For example, all other things being equal, if candidates in year 2 are more motivated than candidates in year 1, then we would expect them to achieve higher marks; if they didn't actually achieve higher marks we would put this down to the test questions being harder in year 2 and would, therefore, set lower cut-scores for the award of each level.

The main body of the present report is a critical evaluation of each of the three main procedures for maintaining national curriculum test standards in Key Stage 2 English and science: pre-test development and statistical scaling; the Angoff procedure; script scrutiny. The first is, essentially, a statistical approach to maintaining standards (which relies on experimental control), while the second and third are principally judgmental methods (which rely on the professional opinion of subject matter experts). All three of the methods are shown to have inherent weaknesses. The major problem with the

statistical approach is that the experimental controls do not always appear to work satisfactorily. The fundamental concern with the judgmental approaches is whether subject matter experts can be assumed to be capable of reaching valid decisions about cut-scores. In fact, it is because there is no one 'foolproof' method for maintaining test standards that national curriculum assessment relies on combining evidence from three.

Perhaps the most significant procedural issue in the national curriculum context is the way in which final cut-score decisions are made. As three different methods are used to arrive at recommendations it is often the case that they conflict. How such conflicts ought to be resolved is an area of academic debate that is not well developed; despite this, it is still not clear that procedures followed in the national curriculum context are as robust as they could be. What does seem to be agreed amongst measurement professionals is that there are rarely, if ever, any 'correct' resolutions when methods for setting standards conflict. However, this does not mean that *any* resolutions will be permissible. On the contrary, each resolution must be, and must be seen to be, *defensible*. Perhaps the most pressing issues that need to be addressed in the national curriculum context are those of *who* should be involved in the decision making process and (when responsibility is taken to lie with a group rather than an individual) *how* individual contributions should be recognised in the final decision.

Principal recommendations from the report are outlined under two alternative strategies. The first strategy, the more conservative one, argues that - while the system is generally defensible - more can be done to improve things. It suggests steps that might be taken in the near future as well as research that will be essential for ensuring the future defensibility of the system. It highlights the need for better documentation of the principles, procedures and practices of standard setting and it addresses the need for a greater public understanding of the system. The second strategy, the less conservative one, argues that a re-focusing of the system would obviate many of the most serious problems currently encountered. The proposal is three-fold: (a) set up a new body to monitor educational standards over long periods of time, perhaps along the lines of the American National Assessment of Educational Performance program; (b) cohort reference national curriculum test results, not awarding levels at all, with the implication that school comparisons be based on the average standard score of pupils; (c) award national curriculum levels to individual pupils purely on the basis of teacher assessment.

Introduction

The present review of procedures for maintaining standards in national curriculum tests is part of a larger review of procedures employed in public tests and examinations.[1] As such, it is intended that this report should feed into a wider discussion of the rationale behind the use of different procedures for aligning assessment standards in the different assessment systems of the UK (e.g., GCE/GCSE, GNVQ, NC tests).

The report focuses upon the maintenance of test standards over time, with particular reference to English and science at Key Stage 2, tests for which the NFER currently has responsibility for developing. (Although the NFER also develops the Key Stage 1 tests for English, Key Stage 1 arrangements will not be discussed.) The report consists of a comment on the quality of current procedures and recommendations for tightening the system.

The report draws on discussions with, and the ideas of; Chris Whetton, Assistant Director at NFER and Head of the Department of Assessment and Measurement; Graham Ruddock, Deputy Head of the Department of Assessment and Measurement and Director of the KS2 National Curriculum Science Tests Development Project; and Marian Sainsbury, Principal Research Officer in the Department of Assessment and Measurement and Director of the KS2 National Curriculum English Tests Development Project.

[1] This project is sponsored by the QCA under the general project title of "Review of models for maintaining and/or monitoring examination/test standards over time" with the specific contract title "Review of model for maintaining and/or monitoring national curriculum tests over time".

1: Models and Procedures

The procedures used to ensure that national curriculum test standards are maintained over time are particularly complex. They require the integration of evidence from many different sources collected at different stages of the test development and awarding process. The principal stages and standard-maintaining procedures are summarised in Appendix 1.

It is important to note that the present system for maintaining standards is a relatively recent development. Before evaluating the effectiveness of current procedures we will briefly reflect upon the ways in which the system has changed since the first national curriculum tests and touch upon some of the underlying conceptual problems that help to explain these changes.

1.1 A brief history of standards in national curriculum assessment

In the original blueprint for the national curriculum, pupils were expected to progress coherently through curricula in a variety of subject areas from Key Stage 1 to Key Stage 4. Each subject was to be sub-divided into profile components which reflected (to some extent) the structure of the subject. Within each profile component, attainment targets were to outline in general the kinds of knowledge, understanding and skills that pupils were expected to achieve by the end of each key stage. Finally statements of attainment within each attainment target were to specify quite precisely the performance standards required of pupils. Different statements of attainment were to be expressed at each of the 10 levels, to describe the performance criteria required for success at that level. The proposed system was considered to be strongly criterion referenced in that standards applied for the award of levels were grounded directly in pupils' performances.

Right from the outset, however, it became clear that the valid determination and maintenance of assessment standards would be a problem. Should a pupil, for example, be awarded level 4 on a profile component if, and only if, she had reached level 4 on *all* of the attainment targets subsumed within it? In such a situation her overall level would be determined by her lowest performance, which seems somewhat unfair. This problem was compounded by the fact that attainment targets differed in the number of statements

of attainment required at each level. How should performance on statements of attainment be aggregated to represent performance on attainment targets? How should performance in attainment targets be aggregated to represent performance on profile components? And how should performance on profile components be aggregated to represent performance in the subject?

One early approach to the first problem was known as the 'n, n-1 rule'. This was initially adopted for setting standards at Key Stage 1. Under this rule, levels for which there were no more than two statements of attainment could only be awarded when criterion performance was reached on all. However, with more than two (i.e., n) statements of attainment, a level could be awarded when criterion was reached on at least n-1. Unfortunately this had the consequence of making it harder to achieve a particular level the more statements of attainment were associated with it (Schagen and Hutchison, 1994).

An early solution to the second problem was to represent performance on the profile component with a particular level if that level had been attained or exceeded in at least half of its attainment targets. This was not without complication, though. A particular oddity reflected the variable number of attainment targets in different profile components: at its most extreme, the level awarded for a profile component with one attainment target would be determined solely by that target.

Similar problems were experienced with the Key Stage 3 written tests. In 1992 mathematics, for example - where each statement of attainment at a particular level was assessed by two unique questions - the original intention was to accept success on at least one of these questions as criterion but to award the level on the test as a whole only if criterion was reached on all statements of attainment. This both proved to be quite harsh and did not take into account performance on the test at higher levels. As a result, a 'rollback' procedure was devised in which performance at higher levels could count towards an award at a lower level. The technique didn't prove to be particularly successful (see Ruddock and Tomlins, 1993; Schagen and Hutchison, 1994).

Implemented at high-speed, the original system for national curriculum assessment soon came under fire from many directions. The main concern of the teaching profession was that it was simply unmanageable: there were too many attainment targets which meant

that too much time was being spent upon assessment - "death by a thousand tick-boxes" as Shorrocks-Taylor (1999, p.13) put it. The main concern of educationalists was that the assessment system would not produce reliable or valid information. Cresswell and Houston (1991), in particular, explained how a system that was founded upon the principle of strong criterion referencing would produce inherently unreliable assessment information; because whether or not a criterion is reached is dependent upon the context in which performance is assessed. Furthermore, as we have just seen, the procedures for determining and maintaining standards simply were not working in practice. With every adaptation, the system appeared to be moving further and further away from its criterion referenced origins.

Teacher discontent reached a high with the test boycotts of 1993 which led directly to a review of the national curriculum and its assessment (Dearing, 1993). The Dearing Review made a number of recommendations, in particular: a reduction of curriculum content in key subjects; a reduction in recording and testing demands; a restriction of the 10 level scale to Key Stages 1 to 3; and the replacement of statements of attainment with level descriptions. This last change was perhaps the most significant in relation to the process of standard setting. Table 1 (abridged from Shorrocks-Taylor, 1999, p.41) illustrates this change.

The switch from statements of attainment to level descriptions indicated explicitly that assessment was no longer expected to be tied directly to 'tick-box' performances but was to be associated more closely with the holistic judgement of pupils' abilities - matching pupils to levels became a process of 'best fit'. This radically reduced the burden of teacher assessment, but it also helped to justify inevitable changes that were occurring in relation to standard setting in tests.

By the time the Key Stage 2 tests were developed (for first national testing in 1995) it seemed inevitable that standards would be set and maintained predominantly on the basis of cut-scores relating to aggregated test marks. This reflected procedures that had evolved for Key Stage 3 which were similar to procedures applied in GCSE and A/AS examinations. The standards were now accepted to be 'criterion related' rather than criterion referenced (Massey, 1995).

Table 1: Statements of attainment (1990) and level description (1995) for the En2 Reading attainment target at level 4.

Level	Statements of attainment (1990)	Level description (1995)
Level 4	a) read aloud expressively, fluently and with increased confidence from a range of familiar literature. b) demonstrate, in talking about a range of stories and poems which they have read, an ability to explore preferences. c) demonstrate, in talking about stories, poems, non-fiction and other texts, that they are developing their abilities to use inference, deduction and previous reading experience. d) find books or magazines in the class or school library by using the classification system, catalogue or database and use appropriate methods of finding information, when pursuing a line of enquiry.	In responding to a range of texts, pupils show understanding of significant ideas, themes, events and characters, beginning to use inferences and deduction. They refer to the text when explaining their views. They locate and use ideas and information.

Standards at Key Stage 2 are still basically determined by applying cut-scores to aggregated mark distributions. No clear consensus has emerged, however, concerning the best approach to take to determining those cut-scores. At present, a variety of methods are employed under the assumption that each is able to contribute important independent evidence. In order to evaluate the effectiveness of these procedures we must interrogate more thoroughly the conceptual framework for standard maintenance that appears, tacitly, to underlie the national curriculum: precisely what is it that ought to be maintained when standards are maintained over time?

1.2 What is it that ought to be maintained when test standards are maintained?

It was originally proposed that national curriculum test results should warrant inferences concerning the 'absolute achievement' of candidates, that is, the award of level X was supposed to reflect a particular standard of performance across a network of domains. As such, the tests were supposed to be criterion referenced, meaning that the standards applied referred to specific performance criteria (i.e., individual P achieved criteria A, B and C). Criterion referencing contrasts most sharply with norm referencing or cohort referencing. Cohort referencing is a better description of what most people refer to as

norm referencing (see Wiliam, 1996). The intention under the cohort referencing model is that if, for example, only the top 5% of a cohort achieve the highest award in year 1, then no more nor less than 5% of the cohort should achieve the highest award in year 2 - no matter how 'good' the year 2 candidates actually were. Cohort referencing simply ensures that the same proportion of pupils achieve each level/grade from one year to the next. In cohort referencing, test levels do not warrant any inferences other than those related to the ranking of candidates' performances within a particular test administration (i.e., individual P performed better than individual Q).

In contrast, the maintenance of test standards within a criterion referenced system amounts to ensuring that the same credit is given for the same performance from one year to the next. This would be the case, for example, if precisely the same test was administered from one year to the next, with the same marking scheme and the same level boundaries. This is the principle underlying the maintenance of standards for coursework in the GCSE and for the Writing paper in Key Stage 2 English.[2]

Unfortunately, however, this approach is problematic when practical considerations mean that only a small sample of learning outcomes are assessed by any one test or examination. In this situation, if the same test were to be applied from one year to the next, then teachers would soon begin to focus upon the assessed outcomes to the inevitable detriment of those that were not assessed. For this reason, national curriculum tests are changed from one year to the next as they sample different learning outcomes. In fact, even when a particular learning outcome is assessed in successive years, the nature of the question through which it is assessed will change.

For practical reasons alone, strong criterion referencing cannot provide the conceptual framework for standard setting in the national curriculum. It is not possible to state, let alone to sample in a test, all possible learning outcomes for a particular key stage. This is before admitting, as we must, that question format can radically affect the difficulty of

---

[2] In fact, it is not clear whether the Writing test can really be said to stay the same from one year to the next, despite the task essentially remaining 'to write' and the marking scheme remaining unchanged. All children are required to write about one specific subject (from a choice of four) each year and the subjects changes from one year to the next. As such, it would not be implausible to assume that the difficulty of the task might change from one year to the next as a function of the subjects chosen. If this were true then, for reasons that are addressed in the following paragraphs, it would not necessarily be appropriate to retain the same cut-scores from one test to the next.

achieving criterion performance (e.g. Foxman *et al.,* 1985; Cresswell and Houston, 1991). On the other hand, it might be argued that norm or cohort referencing would constitute a politically unacceptable alternative conceptual framework. National performance targets mean that national curriculum tests must, in practice, be able to yield information that permits the assessment of progress in performance of the educational system over time. Our dilemma, then, is that while comparability must somehow be grounded in non-arbitrary performance standards, unassailable practical obstacles mean that these cannot be the standards of strong criterion referencing.

Baird *et al.* (2000) describe the conceptual framework that, in practice, seems to underlie standard setting in UK public tests and examinations as weak criterion referencing. Here, level X is not awarded for demonstrating a specific range of level X performances, but for demonstrating a general level X standard of performance. Crucially, weak criterion referencing accepts that 'a general level X standard of performance' may not look the same from one year to the next. Indeed, weak criterion referencing explicitly permits that even if the actual performances observed from candidates in year 2 clearly appeared to be poorer than those of candidates in year 1, it may still be true that candidates in year 2 deserved *more* credit (for example, because the questions that they had to answer were significantly harder).

Weak criterion referencing effectively re-defines comparability in terms of matching (hypothetical) qualities of candidates rather than in terms of matching (actual) qualities of candidates' performances. It suggests that tests be equated not on the basis of the quality of observed performances (at a certain mark) but in terms of the unobserved 'test potential' of candidates that enabled such performances (given the particular assessed learning outcomes and question formats).[3] Test potential is essentially a pragmatic construct which poses the question 'how would these candidates have performed if they had taken test 1 rather than test 2'. The task of comparability under this framework typically amounts to determining marks (although, more usually, just cut-scores) on tests 1 and 2 at which:

Performance on test 1 = Level X test potential = Performance on test 2

[3] Test potential is defined more thoroughly in Appendix 2.

Of course, there can be no independent measure of the hypothetical test potential of any candidate. And this is the problem that plagues any attempt to maintain standards in a weak criterion referenced system. Each of the models described below uses a different method for linking test performances and underlying test potential. Furthermore, each of these methods is based upon different untestable assumptions (see Scharaschkin, 1999; Wainer, 1999), the validities of which must be taken into account when evaluating claims to comparability emerging from the various models.

In summary, the principles of criterion referencing naturally recommend that test standards be maintained through matching identical performances from one year to the next. However, the pragmatics of national curriculum assessment mean that the maintenance of test standards can never be that simple. Conclusions from standard-maintaining procedures are inherently and fundamentally ambiguous. It seems to be for precisely this reason that a multiplicity of models for maintaining national curriculum standards are now used during the test development and awarding process.

The following sub-sections address each of these models in turn, focusing particularly upon: the extent to which the underlying assumptions are appropriate; the extent to which the approaches are reliable; and, more generally, the extent to which the approaches are valid.

1.3 Test development and statistical scaling

1.3.1 Methodology

Both test development and statistical scaling contribute to the maintenance of test standards from one year to the next. Putting it simply, test development provides the solid foundation upon which models for statistical scaling can be built.[4]

Unlike examinations for the GCSE and A/AS, national curriculum tests and test items undergo a stringent series of pre-tests in order to weed out questions that are obviously

[4] In the present report, the term 'scaling' or 'linking' is used instead of 'equating'. This is simply to highlight that the aim, in the national curriculum context, is not to ensure that a mark on test 2 can be expressed in terms of the mark scale of test 1 (as is the traditional intention of equating) but simply that a

going to cause problems. In particular, these are questions that function poorly, either because they do not discriminate between candidates usefully or because they do not appear to measure the same qualities as other items or because they cannot be marked reliably. In addition, pre-testing enables items to be selected that will result in mean marks and standard deviations on the final test being roughly equivalent to those on the same test in previous years. As well as a common *statistical specification* for the to-be-linked tests, valid linkage is also ensured at the test development stage by ensuring a common *content specification*. Thus, test development should attempt to ensure that the to-be-linked tests sample similar content areas, similar types of questions and similar ratios of questions to content areas sampled (see Kolen and Brennan, 1995).

While test development makes a very important contribution, the onus of responsibility for comparability ultimately lies with the methodology for statistical scaling. Scaling requires experiments designed to generate statistical relationships between observed test performances between years. There are a number of ways in which such experiments could be conducted, for example:

1 Require a randomly selected sample of pupils who sit test 2 in year 2 also to sit test 1 (from the previous year) at the same time.[5] Unfortunately, this has the disadvantage that test 1 will already have entered the public domain and will probably have been used to prepare pupils for test 2. As such, it would not give a fair representation of how year 2 pupils would have performed on test 1 had they not been familiar with it. This model is not used within the national curriculum.

2 Require a randomly selected sample of pupils who sit test 1 in year 1 also to sit test 2 (for the next year) at the same time. If this sample of pupils achieves the same mean marks (and standard deviations) in test 1 and test 2, we assume that the tests are comparable and recommend the same level boundaries in year 2 as in year 1. If, on the other hand, the sample of pupils achieves higher mean marks on test 2 than test 1, we assume that test 2 is less difficult than test 1 and recommend

---

equivalence linkage can be established for test 1 and test 2, such that cut-score standards can be 'carried forward'. In fact, it generally only the cut-scores that are linked.

[5] It is assumed that random selection will assure a distribution of ability similar to the population from which it is drawn.

that level boundaries in year 2 be raised accordingly. Of course, 'accordingly' is determined by the statistical scaling procedures and these are discussed below. This is the principal approach taken to investigate comparability for the Reading paper of Key Stage 2 English. Under this approach there are important security issues, as it would be problematic if test 2 entered the public domain prior to the live test date.

3 Require a randomly selected sample of pupils who sit test 1 in year 1 also to sit an additional 'anchor' test; then require a randomly selected sample of pupils who sit test 2 in year 2 also to sit the same additional anchor test. The anchor test is intended to measure the same quality as is measured by the live tests of year 1 and year 2. It acts as a common yardstick, or reference, against which test 1 and test 2 can be calibrated - it anchors the standards. As such, if the sample of pupils in year 2 achieved higher mean marks on the anchor than pupils in year 1, then we would expect higher mean marks on test 2 as well. If the same mean marks were observed on test 2 as had been observed on test 1, we would assume that test 2 was more difficult than test 1 and would recommend that level boundaries in year 2 be lowered accordingly. This approach is adopted as an additional insight into comparability for the Reading paper of Key Stage 2 English.

In model 2 above, we assume that the average test potential of pupils taking test 1 is the same as the average test potential of pupils taking test 2. This seems to be a reasonable assumption as they are the same pupils. Then, as both tests are assumed to be equally valid measures of the same construct, we assume that any difference in mean mark is attributable to the difficulty of the paper. If any effect of difficulty is observed then the mark distributions will need to be scaled for equivalence. Scaling simply means projecting marks (actually just the cut-score marks) on test 1 to marks on test 2 which represent the same level of test potential. Through this process of mark scaling, we project cut-scores directly from test 1 to test 2. Scaling is not actually done on the basis of individual pupils' scores (as an individual pupil may not necessarily demonstrate precisely the same test potential from one administration to the next, for one or more of many possible reasons). Instead it is done by comparing the overall distributions of marks on each test (assuming that chance variations across pupils average out over a large sample).

Scaling marks across papers can be achieved using either a linear or an equipercentile model. Linear scaling matches marks on the basis of the relative distance of marks from the mean mark on each test. Equipercentile scaling matches marks on the basis of the proportions of pupils at each mark. In linear scaling, if the level 5 cut-score lies (say) ¾ of a standard deviation above the mean mark in test 1, then the test 2 level 5 cut-score is deemed to correspond to the mark that lies ¾ of a standard deviation above the test 2 mean. In equipercentile scaling, if (say) 70% of pupils achieve below the level 5 cut-score on test 1, then the test 2 level 5 cut-score is deemed to correspond to the mark below which 70% of pupils achieve on test 2.

The linear method works well when two tests have similar mark distributions; it also tends to be more useful than the equipercentile method with small sample sizes. When the mark distributions for two tests are differently shaped (for example, if one is skewed) then the equipercentile scaling method is more appropriate (e.g. Green, 1995). The two techniques tend to agree most at the middle of the mark range and least at the extremes. When the techniques do disagree, if reasonable sample sizes have been used, the equipercentile prediction tends to be preferred to the linear one.

In model 3, test 1 cannot be equated directly with test 2; instead, both have to be equated against the common anchor test. Assuming that test 1 and the anchor test are equally valid measures of the same construct, we are able to use either linear or equipercentile scaling to represent the cut-scores on test 1 in terms of the anchor test mark scale. Likewise, assuming that test 2 and the anchor test are equally valid and reliable measures of the same construct, we can use either linear or equipercentile scaling to represent the projected cut-scores on the anchor test in terms of the test 2 mark scale. This assumes that the amount of test potential required for success in one administration of the anchor test is the same as that required in the next. This seems to be a reasonable assumption as the same test is used.

1.3.2 Conceptual rationale

As indicated earlier, the conceptual framework for comparability defined in terms of test potential typically amounts to the question 'how would these candidates have performed if they had taken test 1 rather than test 2'. In principle, the methods described above represent either direct, or nearly direct, experimental tests of this question and, therefore, are well fitted to the purpose of maintaining national curriculum test standards over time.

However, these techniques all make one very important assumption: that the linked tests measure the same qualities. If test 1 does not measure the same quality as test 2 (or the anchor test) then the logic of scaling is compromised. This has implications for the degree of rigour that is required during test construction. Moreover, it has implications for the stability of the curriculum for which each test is constructed: curriculum change immediately compromises the logic of scaling prior and subsequent tests. This is embodied in the construct of test potential which is always defined in relation to a specific curriculum.

1.3.3 Technical issues

Feuer *et al.* (1998) note that it is possible to establish valid links between two tests that meet certain criteria, namely, when the tests are:

1. Created to identical specifications
2. Highly similar in content emphasis, difficulty and format
3. Equally reliable
4. Expected to be administered under the same conditions

If these criteria have been met then we will have two tests which measure the same construct in the same way. In fact, if two tests really did fulfil all of these criteria then little in the way of scaling will actually be required. In practice, this is unlikely. The reason that we need to scale national curriculum test results is, basically, that we cannot guarantee that the questions will be of equivalent difficulty between years. Most of the

problems that we face in accommodating this factor are caused by violations of the other assumptions.[6]

When discussing national curriculum tests, we refer to a particular problem of scaling as the 'pre-test effect'; the present report will define this as the tendency for different (average) levels of test potential to be exhibited between the two tests that are being linked when the scaling is conducted.[7] The fundamental assumption of scaling methodology is that the (average) level of test potential is controlled across to-be-linked tests; this means that pre-test effects are serious. It appears that pre-test effects occur for many of the national curriculum tests; these effects are suggested by mean differences in performance between the pre-test sample and the final test cohort (i.e., different candidates, on the same test, separated by a year). The apparent pre-test effects in Key Stage 2 Reading and Writing were of the order of 2 to 3 marks for 1997, 1998 and 1999 (from a mark total of 50 in each year and for each paper). Indeed, Whetton (2000) suggests that these pre-test effects may be increasing over time.

To the extent that pre-test effects can be attributed to differences in the test potential of pupils taking the to-be-linked tests, two possible explanations have been suggested:

1. Test potential for the pre-test is lower than test potential for the concurrent live test (even amongst the same group of pupils), because pupils are more motivated to achieve well in the latter than the former.

2. Test potential for the pre-test is lower than test potential for the concurrent live test (even amongst the same group of pupils), because the live test tends to be taken slightly later in the school year, presenting an opportunity for test practice as well as meaning more time for curriculum coverage, revision or 'cramming'.

Both of these explanations are, essentially, violations of the fourth criterion identified by Feuer *et al.* (1998), as the tests are not being administered under the same conditions.

---

[6] Of course, it is important to remember that all tests have a degree of unreliability inherent in them, even in the most favourable of situations. Even if you used the same test twice, with the same 1000 pupils, you might not reach the precisely the same cut scores after equating (Quinlan and Scharaschkin, 1999).

[7] This definition is adopted to distinguish effects that result directly from weaknesses in the methodology of scaling and those that arise for other reasons.

The inference that pre-test effects are present (from the evidence of a mean mark difference between pre-test in year 1 and final test in year 2) is based on two prior assumptions: (i) that that the pre-test sample is selected at random; and, consequently, (ii) that there is no mean difference in test potential between the pre-test sample in year 1 and the final test cohort in year 2. Of course, if there *were* overall differences in test potential (due, perhaps, to the effect of a national strategy), then an average difference in performance between the pre-test and final test would not be directly indicative of pre-test effects.

Similarly, marking differences between pre-test and final test would confound estimates of genuine pre-test effects. That is, if markers were not rewarding similar performances in the same manner, between pre-test and final test, then differences in performance would be indicated when none actually existed. In fact, there are good reasons to suggest that marking effects may occur; in particular, the fact that final test scripts undergo borderline re-marking. During borderline re-marking, scripts from those pupils directly below each level boundary are re-marked. However, as only mark increases are recorded this will act to inflate the mean mark for the final test cohort artificially. Quinlan and Scharaschkin (1999) suggest another potential marking issue. As the pre-test is marked by a small team of experienced markers, while the final test is marked by a larger number of less experienced markers, there may be more chance of a non-random interpretation differences between the two. A final marking issue is the risk of interpretation effects being even more pronounced when marking schemes change between pre-test and final test.

This raises another issue that confounds the interpretation of pre-test effects. Occasionally, it is not only the mark scheme that changes from pre-test to final, but the actual test. Question order may be altered, slight wording changes may occur, or questions may even be removed or replaced. If questions change then effects upon the final test results can be estimated to a certain degree (as the questions will have been pre-tested in a different format). However, untestable assumptions must be made; for example, that the questions will be of the same facility in the new test. It has been shown that such assumptions do not necessarily hold and the facility of a question can differ even as a function of its order in a test (e.g. Goldstein, 1996). (Although, it must be said, much

test theory and most test development depends on the principle of independent functioning of items.)

Returning now to the putative pre-test effects - the potential weaknesses of the methodology used for linking national curriculum tests - evidence pertinent to the two explanations presented above was discussed by Whetton (2000). The fact that pre-test effects are largely absent from Key Stage 3 mathematics raises the question of whether procedural differences can explain the problem. In fact, a notable aspect of the procedures for Key Stage 3 mathematics is that the pre-testing is carried out shortly *after* the previous year's live test, rather than before (pre-testing appears generally to occur before live testing). This raises the interesting possibility that inflated performance on the live test may, to some extent, be due to practice or due to the additional curriculum coverage, revision or 'cramming' that is possible when pre-tests are taken first. The lack of pre-test effect for Key Stage 3 mathematics also seems to argue against a differential motivation effect between pre-test and live test (however, it could simply be that pupils are not more motivated in live tests than pre-tests at Key Stage 3, if these are perceived as 'low stakes'). Further research is needed to clarify the nature of these effects.

Moving to more conceptual issues, it is important to the logic of scaling that the to-be-linked tests measure the same thing. To the extent that the qualities assessed by two tests do not, problems can arise that are hard to resolve (Feuer *et al.*, 1998; Goldstein, 1986; Goldstein, 1996). One such problem is that different scaling functions may result when computed for different sub-groups of the population separately. For example, a boy and a girl with the same score on test 1 might be scaled to different scores on test 2. This might happen, for example, if girls and boys were equally motivated for test 2, while girls were more motivated than boys for test 1 (or, similarly, if test 1 measured an aspect of the curriculum that girls tended to excel at). The point is not that one or other of the sub-group scaling functions is 'right', nor that the overall scaling function is 'wrong', it is simply that the concept of scaling is rendered somewhat meaningless in such situations. If test 1 is measuring a different quality from test 2 then we are left grappling with an unresolvable problem of what comparability could possibly mean in such a situation. The problem is not technical, it is conceptual. An alternative way of framing the problem would be to express it in terms of test potential. Differing scaling functions for identifiable sub-groups mean that the test potential of candidates in test 1 will differ from

the test potential of candidates in test 2 (even though they are the same candidates). This undermines the basic logic of scaling methodology which is to control for differences in test potential. Note that this problem can be exacerbated by the use of an anchor test when scaling functions change over time. Likewise, scaling functions can be different between sub-groups of different ability, causing problems for scaling across tiers.

> "Equating works primarily because of the care that goes into ensuring that the tests measure the same construct. ... A slight change in topic emphasis from year to year can change the amalgam slightly, which can mean that this year's test isn't measuring quite the same thing as last year's test, which renders the equating suspect." (Green, 1995, p.14)

Scaling is rendered suspect when curricula change, or when test formats are altered significantly. This is a particular problem for national curriculum assessment. For example, the 1999 Key Stage 3 English test included a new Shakespeare play, which meant that the test development agency had to "make allowance" for this (Quinlan and Scharaschkin, 1999). Exactly what making allowance should amount to is a problem that has no simple solution. Curricula inevitably change in other subjects as well; for example, the introduction of more 'number crunching' or more algebra into the mathematics curriculum, or (as actually happened) the addition of a new mental test. Any changes render scaling suspect and force untestable, and sometimes somewhat arbitrary, assumptions to be made. The Reading paper of Key Stage 2 English appears to be particularly vulnerable to this challenge as there has been a policy of changing the type of text used from one year to the next.

Finally, a word on the anchor test. While repetition of the same test from one year to the next appears intuitively attractive as a method for monitoring standards, it carries with it some distinctive problems. In particular, if the anchor test is shorter than the final test this will reduce its reliability (a violation of Feuer *et al.*'s third criterion). Moreover, this is also likely to reduce its validity, as fewer curriculum areas may be sampled in a short test. Indeed, when pupils also take a reduced pre-test this weakens the validity of scaling even further. There is also the problem that, if the curriculum changes, the anchor test will become invalid (see above discussion). Finally, there is the problem that the reduced mark scale of the anchor test inevitably places a practical limit upon the validity of marks

projected upon a larger scale. The more imprecise the matching of boundaries becomes the more potential for 'standards drift' exists. In short, the use of the anchor test in Key Stage 2 English presents a weak link in the comparability chain.

1.4 Angoff procedure

1.4.1 Methodology

The assumption behind Angoff's standard setting procedure (e.g. Angoff, 1971) is that experienced teachers are able to use mental models of pupils' abilities to judge how difficult borderline pupils will find individual questions.[8] Having thus determined the difficulty of individual questions, these estimates can be aggregated to determine the difficulty of the test as a whole. This leads directly to estimates of appropriate cut-scores. The procedure can be illustrated with respect to English at Key Stage 2 in 1999.

The Angoff level setting meeting for Key Stage 2 in 1999 was held over a period of two days during November 1998. The intention of the meeting was to generate three cut-scores on the Reading test: level 3, level 4 and level 5. Thirteen participants were selected from across England with the requirement that they be experienced Year 6 teachers with expertise in English and/or assessment. They represented teachers from primary and junior schools of varying sizes from a variety of types of area.

During the training phase of the exercise, the teachers took part in group discussion training sessions in which they formulated descriptions of borderline pupils (at levels 3, 4 and 5) and applied these to the kind of single and multiple mark items that they were to evaluate subsequently. For the purpose of the study a borderline pupil was defined as a pupil who, according to teacher assessment, would be judged as *just* attaining the level.

During the first phase, each of the teachers, working individually, was required to estimate the proportion of borderline level 3, 4 and 5 pupils that they considered would be successful on each actual test item. This was phrased in terms of imagining 100 borderline pupils and estimating how many would be successful on each question. On multiple mark items, the participants divided their 100% between each of the marks (e.g.,

[8] Borderline is defined as the lowest mark at which a certain level is awarded.

10% would achieve 0 marks; 30% 1 mark; 50% 2 marks; and 10% 3 marks). From these initial estimates, cut-scores for each paper were derived for each teacher, at each judgmental boundary, by aggregating the following product across items: the predicted proportion of successful level X pupils (from 0 to 1) multiplied by the total number of marks attained on each question.

During the second phase, teachers were given feedback concerning how their decisions compared with those of the rest of the group. Statistical information was also provided (from the second pre-test) concerning the facility of each of the items and the accompanying distribution of marks for the test as a whole. Reflecting upon this information, but still working individually, teachers were given the opportunity to revise their original estimates.

During the final phase, a group discussion was convened which focused particularly on differences in estimates between group members (particularly large ones) and the reasons for these differences. Once more, the teachers were allowed to reconsider their judgements in the light of this new information. From teachers' final item-level estimates, their final cut-scores were computed; these were averaged across teachers to provide the cut-scores recommended by the group as a whole.

1.4.2 Conceptual rationale

The use of the Angoff procedure is a recent development in the UK, although it has been used far more widely in the USA. Its use for setting national curriculum standards was pioneered in Northern Ireland in 1993 (Morrison *et al.*, 1994) and it has been adopted throughout the UK since then. Angoff's procedure has been interpreted in numerous ways, but the approach adopted in the UK is based upon the 'eclectic' model recommended by Berk (1986). This model incorporates the features that Berk saw as important for ensuring maximum reliability.

The basic assumption behind the Angoff procedure is that teachers are able accurately to estimate the probability that a 'minimally competent' borderline pupil will achieve success on a particular question. In terms of the weak criterion referencing model described earlier, it presumes that teachers are able to manipulate the following in order to determine their probability judgements: mental models of pupils with threshold test

potential and judgements concerning the test potential required for successful performance on individual questions. Two conceptual problems immediately arise. First, is it reasonable to assume that teachers can form valid mental models of pupils with threshold test potential? Second, is it reasonable to assume that teachers can accurately determine the level of test potential required for successful performance on individual questions?

First, is it safe to assume that even experienced teachers possess valid internal representations of borderline standards? In a genuinely criterion referenced examination, it might be argued that criteria would not only be written down in black and white for all teachers to internalise but would also have been grounded in the practical assessments of teachers in the first place. However, the national curriculum is not a genuinely criterion referenced system and it is debatable whether level descriptions are sufficiently precise to allow teachers to use them as the basis for constructing representations of minimally competent pupils. This is particularly so as level descriptions are defined in terms of typical pupils at a level, not borderline pupils. It is important to spell out the problem carefully. The issue is not whether teachers have sufficient professional knowledge to be able to use and apply level descriptions in a generally satisfactory manner; no doubt they can, and frequently do, in the everyday classroom context (see Sainsbury and Sizmur, 1998, for a discussion of the conceptual framework underlying teacher assessment in the national curriculum). The question is more specific: (a) when even level descriptions are defined in terms of the performance of *typical* level X pupils, how reasonable is it to expect teachers to be able to generate valid representations of pupils who fall into the grey area *between* levels? (b) moreover, how reasonable is this when the degree of precision required is such that teachers are required to represent, not simply the grey area, but a specific point within that grey area - the borderline (a borderline pupil represents the test potential relating to a single mark on a mark distribution - neither one mark more nor one mark less)?

Cizek's answer to the problem of construct confusion was as follows:

> "in order for a standard setting method to be implemented properly, clear definitions of key constructs must be developed a priori. In practice, participants will usually need to make repeated references to a formal, written summarization

of the attributes and performance indicators that define the construct of interest in order to maintain conceptual fidelity and to implement the procedure faithfully." (Cizek, 1996, p.15)

Unfortunately, even ignoring the fact that there are no written statements concerning national curriculum borderline pupils, Cizek's guidelines are compromised by their strongly criterion referenced assumptions. We cannot assume that 'attributes and performance indicators' alone are capable of *defining* our construct of interest. This is not to deny that clear task specification might be of assistance in *easing* conceptual confusion - as long, of course, as the criteria specified demonstrably related to standards that had been applied in previous years! What should be done, for instance, if specifications (in level descriptions) of what borderline pupils *ought* to be able to achieve conflicted with evidence (from previous examinations) of what borderline pupils *were* able to achieve? Perhaps, in the context of maintaining standards, it might be safer to require that teachers' understandings of minimally competent pupils be developed in practice on the basis of feedback from standards set in previous tests? This would have implications for the use of the Angoff procedure in the first few years of testing in a curriculum area.

Turning to the second question, is it safe to assume that even experienced teachers are able to judge question difficulty accurately? There seems to be a major conceptual dilemma here: if the task of perceiving differences in the difficulty of apparently similar questions was not quintessentially problematic, then the Angoff procedure would not be needed at all: test writers would simply write questions of identical difficulty from one test to the next and level boundaries would remain unchanged. If teachers cannot be relied upon accurately to account for question difficulty then this calls into question the appropriateness of the procedure (see also Massey, 1995).

In fact, it will be noticed that the national curriculum version of the Angoff procedure - following the method described in Morrison *et al.* (1994) - ensures that scrutineers are allowed to revise their initial judgements on the basis of statistical information concerning the difficulty of individual questions. Morrison *et al.* (1994) demonstrate how this tends to bring teacher judgements of relative question difficulty into line with statistical estimates. This effect might be taken to suggest that teachers are not necessarily confident that they can determine question difficulty with more precision than statistical

measures. More importantly, though, we might question the validity of asking teachers to integrate statistical information at all. A particular strength of the Angoff procedure is its face validity: the determination of standards by professional judgement. It is not clear the extent to which this is reduced when professional judgement is modified by the impact of statistical information. To what extent can the final judgements really be considered to represent *teachers'* standards?

1.4.3 Technical issues

There is very little UK research into the reliability or validity of the Angoff procedure for setting educational standards.[9] However, a considerable body of American research has built up over the past two decades. Although a detailed review is beyond the scope of the present report, some indication can be given of the main issues that have arisen. Unfortunately, there seems to be no clear consensus view from the American literature. While few, if any, would accept that the standards embodied in the Angoff procedure are entirely trustworthy, opinion is divided concerning the seriousness of its limitations. As noted by Dwyer (1996), some of the most distinguished measurement experts are so unimpressed by available methods that they recommend that standards *should not be set at all* unless absolutely necessary. Of course, standards do have to be set for many tests, which is why many researchers have adopted the approach of seeking out the best method from what may (or may not) be considered a bad bunch.

Berk (1986) took exactly this approach, providing a 'consumer's guide' to standard setting procedures. He reviewed 23 methods for setting cut-scores using techniques that relied upon judgmental decision making and he evaluated each in terms of the extent to which they fulfilled a number of technical and practical criteria. Technical criteria concerned issues such as reliability and validity; while practical criteria concerned whether the procedure was easy to implement and understand. Berk also made an important distinction between purely judgmental methods such as the original Angoff and judgmental-empirical methods such as the iterative Angoff. The latter differs from the former as participants are allowed to revise initial recommendations in the light of actual

[9] There are numerous versions of the Angoff procedure (see Berk, 1986). The report will use the generic 'Angoff' when greater precision is unnecessary.

performance data (particularly item facility indices). The procedures used in national curriculum assessment are, therefore, judgmental-empirical.

Berk concluded that the Angoff procedure provided the best balance between technical adequacy and practicability, from amongst the judgmental approaches. However, he did note two problems: firstly, that many judges have difficulty defining students who are minimally competent; and, secondly, that certain test items are more difficult to rate accurately than others.[10]

While the judgmental approaches were deemed simpler and, therefore, more practical than the judgmental-empirical methods, the latter won out in terms of technical adequacy. In actual fact, within the latter category, the iterative Angoff procedure lost out overall to the informed judgement method (Popham, 1981). However, as already mentioned, when finally recommending a model technique for setting standards for educational certification, Berk did borrow heavily from the iterative Angoff. As such, it would seem that he generally rated this approach above the others (for the purposes that we are interested in).

Morrison *et al.* (1994) reported the results of a study that translated Berk's recommended model into the national curriculum context. They illustrated how the iterative procedure was particularly effective in ensuring that teachers' estimates of question difficulty better reflected the difficulty ranking given by item facility indices, an indication of enhanced validity, perhaps. Moreover, the variability of teachers' judgements also decreased notably from session to session after statistical feedback and group discussion. They concluded that "reliability of recommended standards increases both as a consequence of receiving normative data and of discussion" (Morrison *et al.,* 1994, p.181). Thus, the use of the iterative Angoff procedure is defensible in terms of its proven ability to enhance judgmental reliability to a greater extent than many other judgement-based approaches. However, it important to note that enhanced reliability is of little comfort if all that is ensured is a more systematic bias. We need to be confident that participants are converging on the appropriate decisions. Whether Morrison *et al.* (1994) convincingly

---

[10] "The subjectivity of the item content decisions used to arrive at the performance standard for most of the methods can be expressed as follows: Judges have the sense that they are "pulling the probabilities from thin air" (Shepard, 1980a, p.453)." (Berk, 1986, p.147)

demonstrated this is open to debate, as is the issue of whether their statistical estimates of reliability and validity are appropriate (Massey, 1995).

It is worth returning to the two questions raised earlier, the first of which asked whether teachers are really able to conceive of a borderline pupil. A study by Sizmur (1997) raised a note of caution in this respect. In a study of standard setting for reading tests for 7-year-olds, Sizmur concluded that the cut-scores from the iterative Angoff procedure were too low (a conclusion reached through triangulation with other sources of information). He suggested that this could be at least partly explained by a lack of familiarity on behalf of the teachers with how pupils would interact with the type of test used. Without sufficient familiarity with the test type, teachers were not able to picture the likely performance of a borderline candidate.

Teachers involved in Angoff exercises for national curriculum tests would undoubtedly have more familiarity with the testing formats and with the ways in which their pupils would respond to them. Indeed, it is likely that this 'guild knowledge' (Cresswell, 1996) will increase with every year of national curriculum assessment. However, whether even they are able to form accurate mental pictures of borderline pupils at various levels is a research question that remains to be addressed. Research within the national curriculum context would seem particularly important as the rationale used here differs from that typically reported in the American literature. In the UK, judges are supposed to be *maintaining* a pre-existing standard rather than *setting* a new standard; this means that the standard to be carried forward has a very precise specification in relation to discrete marks on a previous year's test.[11]

The second question that we need to reconsider is whether teachers are able to assess the absolute degree of difficulty that would be experienced by borderline pupils attempting test items. The study by Morrison *et al.* (1994) suggested that, on average at least, teachers were not bad at estimating the relative difficulty of questions (as indicated by the

[11] Morrison *et al.*'s (1994) recommendation that teachers "identify such a minimally acceptable pupil (or pupils) from among the young people they teach" assumes that each participant teaches at least one genuine borderline pupil. Even this assumption is dubious in light of the specificity required to maintain test standards within the national curriculum. A predictable reply might be that errors in mentally representing borderline pupils would be distributed randomly around the 'genuine' borderline pupil (and so cancel out). Perhaps. If nothing else, though, this raises issues of the size of the sample of teachers.

correlation between average probability estimates and actual item facilities). However, the reason for using teachers is not to assess relative difficulty but absolute difficulty - absolute difficulty for borderline pupils. Once again, the jury is still out concerning whether teachers are able to perform this function effectively.

Impara and Plake (1998), following Bejar (1983), believe that they cannot. They conducted a study in which teachers first defined 'D/F borderline' with respect to their own students by predicting their grades in a sixth-grade benchmark science test which had been in use for four years. Across 26 teachers, borderline students were deemed to be the 95 who were predicted to achieve either D or F (where F was the 'failing' grade, directly below D). The same teachers then followed an iterative Angoff procedure to rate the facility of items in the test for "those students in your class who are just barely passing the class (the borderline D/F student)". Their results showed that, for half of the teachers involved in the study, the *average* difference between estimated item facility (for a borderline D/F pupil) and actual item facility (across the 95 D/F pupils) was greater than or equal to 0.25.[12] Putting it simply, while the teachers were reasonably accurate in determining the relative difficulty of questions, they were not accurate in determining the absolute difficulty of questions. Importantly, this appeared not to be a random effect, but a consistent effect of under-estimating the apparent facility of items (i.e., they predicted lower success rates than were observed).

Unfortunately, there is a technical hitch in this study: if there happened to have been more pupils predicted D than F, then we might have expected teachers' ratings to under-estimate the facility of items for genuine borderline pupils (when compared with the actual facilities of items for a group in which the average ability would clearly have been higher than the borderline). Sadly, information is not available to shed light upon this matter (Impara, personal communication, 8 March 2000). However, this is an important and suggestive study that would appear worthy of replication in the national curriculum context.

---

[12] This is 0.25 from a scale of 0.00 to 1.00, i.e., estimates of the proportions of successful borderline pupils erred, on average, by 25%.

Finally, it is worth mentioning issues related to the sampling of participants for an Angoff exercise. While not always statistically significant, Morrison *et al.* (1994) noted a tendency for absolute estimates of difficulty to differ by type of rater. Grammar school teachers tended to rate questions as easier than Secondary or Primary teachers and tended to be less resistant to change their ratings in the light of performance data. Similar effects have also been noted in Key Stage 3 mathematics (Ruddock, personal communication, 17 March 2000). While we cannot say that one type of rater is necessarily more accurate than another, if genuine rater type effects do occur then there are particular sampling issues that need to be addressed. When the Angoff exercise is conducted with relatively small committees there ought to be a very good rationale for the choice of participants. It is worth bearing in mind that phase 1 and phase 2 of the iterative procedure are perfectly amenable to a large scale survey approach. Perhaps a two-stage iterative approach with large numbers of teachers might prove more valid and reliable than a three-stage iterative approach with a chance for group discussion but fewer participants? Once more, this is an issue that deserves further research.

1.5 Script scrutiny

1.5.1 Methodology

The assumption behind script scrutiny is that subject specialists are able to use mental models of pupils' underlying abilities to adjust for the differing difficulties of test versions. Live scripts, each at a specified mark, are individually scrutinised in order to determine their standard in relation to a historical benchmark. The historical standard is exemplified in scripts from pupils deemed to have been at the borderline of each level on the same test paper in the previous year.

The script scrutinies, for each subject at Key Stages 2 and 3, constitute the initial phase of the Level Confirmation Exercise and take place shortly after the live test administrations. Committees are formed from subject specialists including the test's Lead Chief Marker, marking agency Chief Markers, Deputy Chief Markers and others. The specialists begin their exercise by scrutinising a sample of scripts from the previous 3 years' tests at the

relevant level boundary of the paper in question.[13] They continue with a discussion of the "qualities of those scripts, the characteristics of performance they embodied, as well as any concerns that they might hold about the marking of those scripts" (Quinlan and Scharaschkin, 1999). In 1999, the Key Stage 2 English scrutineers considered the Reading and Writing components separately (level cut-scores for each boundary are aggregated across Reading and Writing to form the overall English test boundaries).

Having familiarised themselves with the historical standard, the subject specialists scrutinise a small number of packs of scripts from the live test, where each pack corresponds to scripts at one of a range of marks around the likely threshold. In Key Stage 2 English in 1999, scrutineers judged 12 scripts at each mark (from a 5 mark range around the likely boundary) to determine cut-scores for each of levels 3, 4 and 5 for Reading and Writing, respectively. The sample scripts had been photocopied, during which names and mark totals had been eliminated so that - in theory - scrutineers should not have been able to tell which packs represented the higher marks within the range and which represented the lower marks.

Working independently, the specialists judged which packs contained scripts 'below the historical standard', 'equivalent to the historical standard' or 'above the historical standard'. These judgements were collated on a flip chart to represent the level of agreement between individuals. Group discussion was then employed to reach consensus concerning which packs most accurately matched the quality of performance of borderline pupils in previous years. The marks awarded to scripts in these chosen packs determined the final boundary recommendations.

This procedure is very similar to that which is encapsulated in the Code of Practice for GCSE and A/AS examinations. However, at GCSE and A/AS, subject specialists are explicitly given additional evidence relating to the component in question (in particular, statistical evidence concerning the implications of different cut-scores for the proportion of pupils that would be expected to reach each grade).

---

[13] These archive scripts are sent to participants prior to the meetings and, by the committee stage, the scrutineers should have read them and formed mental models of the standards embodied at each boundary.

1.5.2 Conceptual rationale

The use of script scrutiny to inform (or determine) comparability decisions has a long history within the UK public examining system at 16+ and 18+. Indeed, script scrutiny meetings are the central focus of standard setting in the GCSE and A/AS and they have been for many years. As Christie and Forrest (1980, p.21) explain: "The requirements that boards make of their senior examiners are such that they implicitly attribute to them the ability to spot a borderline script at twenty paces." Unfortunately, the theoretical underpinnings of this practical requirement have hardly ever been made explicit. Similar problems arise in this context as arose with the Angoff procedure: what exactly does a borderline pupil look like and is it likely that an examiner would have the ability to recognise one?

Recent work by Cresswell, Baird and Newton (e.g. Cresswell, 1996; Baird *et al.*, 2000; Newton, 1997a, present report) has attempted to clarify the situation by specifying more tightly the definitional framework for comparability within UK public examinations and the task required of script scrutiny panels. This is how Baird *et al.* (2000) frame the problem:

> "In the simplest situation, two examination papers would sample precisely the same curriculum areas (using slightly different questions). This would enable awarders to make direct comparisons of performance, between the two examinations, curriculum area by curriculum area. In fact, this situation is typically not observed in public examinations and the two examination papers will generally sample somewhat different curriculum areas (again, using slightly different questions). This makes the task for the awarders considerably more difficult as they now have to be able to take into account how candidates for each of the examinations *might have* performed in curriculum areas that were not sampled. This, presumably, requires some kind of extrapolation from performance in shared curriculum areas to performance in curriculum areas that are not shared. It may be that, as teachers, awarders can carry out this task using their experience of personally assessing students throughout the year across the full range of curriculum areas. This, at least, may be the case when the curriculum remains fairly constant from one examination to the next. However, the task

becomes significantly more difficult when the curriculum changes. In the most extreme situations, when experienced awarders have been asked to judge the comparability of examinations based upon syllabuses that have changed radically over time, even many of them have been forced to conclude that the task is, essentially, impossible (e.g. Christie and Forrest, 1981; SCAA, 1996)."

In terms of test potential, the task of the scrutineer would be to extrapolate from observed performance at a boundary on the previous year's test to the underlying test potential of pupils at that mark. She would then have to project this test potential onto the present test to construct the performance that might have been expected from last year's borderline pupils had they taken this year's test. Putting it simply, the basic assumption of the script scrutiny method is that scrutineers are able to 'see through' question/test difficulty. However, as discussed in relation to the Angoff procedure, if the task of perceiving differences in the difficulty of apparently similar questions was not quintessentially problematic then script scrutiny panels would not be needed at all: test writers would simply write questions of identical difficulty from one test to the next and level boundaries would remain unchanged. In principle, then, the conceptual rationale for script scrutiny is, once more, slightly questionable.

1.5.3 Technical issues

Although script scrutiny has been used to inform level setting for a number of years there has been little, if any, research into the appropriateness of this procedure in the national curriculum context. Indeed, while there have been many studies which have used script scrutiny panels to monitor comparability at GCSE and GCE A/AS (e.g. Bardell *et al.*, 1978; Christie and Forrest, 1980; Forrest and Shoesmith, 1985; Fowles, 1995; Quinlan, 1995), there have been few which have attempted to assess directly their reliability or validity. The important exception is work that has been conducted by Cresswell and colleagues at the Associated Examining Board (e.g. Cresswell, 1997; Cresswell, 2000; Scharaschkin and Baird, forthcoming). The conclusions of this research are relatively pessimistic concerning the validity of subject matter expert consensus as the sole arbiter of comparability.

Cresswell (1997) investigated the validity of awarding decisions, for a variety of medium to large entry A-level subjects, made at a time when grade boundaries were fixed purely on the basis of subject matter experts' judgements. Considering the outcomes of awarding meetings in two adjacent years, he began by determining the degree of change in the proportions of candidates achieving grades A, B and E for each subject. Across 38 subjects, the degree of change in cumulative percentages of pupils achieving each grade ranged from: 0% to 13.2% at grade A (>=5% in 3 subjects); 0.1% to 15.7% at grade B (>5% in 13 subjects); and 0% to 15.3% at grade E (>5% in 14 subjects). Cresswell's intention was to consider whether, on the balance of probability, the profile of changes observed across subjects was rationally defensible.

To defend changes in the cumulative percentages of pupils at a specific grade, one of three explanations might be given:

1. As a whole, candidates' performances had changed from year 1 to year 2 (e.g., candidates were generally better in year 2);

2. The balance of identifiable sub-groups within an examination cohort had changed from year 1 to year 2, where the sub-groups tend to display different performance characteristics (e.g., there were proportionally more female candidates in year 2 and female candidates tend to perform better);

3. New (or missing) candidates in year 2 displayed different performance characteristics from the year 1 examination cohort (e.g., the examination had attracted a large number of new entries from independent selective schools).

Dismissing explanation 3 was relatively simple. The degree of change in number of candidates between years was not sufficient to explain the degree of change in performance profiles. Year-on-year changes in the total numbers of candidates for each subject tended to be small in comparison with existing candidate numbers.

Explanation 2 required more evaluative work. For each subject, Cresswell re-weighted the proportions of candidates from identifiable sub-groups at grade X in year 1, by the proportion of the year 2 cohort that they represented. Then the predicted changes, due to differential representation of sub-groups in year 2, were compared with the actual changes

observed. In fact, not only were the observed changes considerably larger than the predicted changes, they were not even correlated with them.

Explanation 1 was the most complicated to evaluate. Cresswell began by eliminating one obvious explanation: that the observed changes were due to genuine performance differences, but ones due to chance alone. Simple statistical modelling demonstrated that the pattern of changes across subjects did not reflect that which could be put down to chance alone. If not due to chance, potential alternative explanations had to be considered, the first of which was that a factor common to the entire year group had led to changes in performance across all subjects. In fact, this was easy to rule out because, for some subjects, performance appeared to increase while, for others, it appeared to decrease. The obvious reply, though, was that whatever caused the increases or decreases was subject-specific, for example, motivational changes in a subject area from one year to the next.

Cresswell's counter was two-fold. Firstly, he examined performance changes within subjects for sub-groups of candidates from different types of institutions: schools, further education colleges and from overseas. This revealed a strong tendency across subjects for changes in performance from one year to the next to be in the same direction for all three sub-groups. The implication is that if a subject-specific factor had been at work, it had had a similar impact upon students from a range of different backgrounds. It is not clear how or why a factor such as general motivation could have had such a uniform impact, particularly when considering pupils from different countries. Secondly, Cresswell investigated changes observed between year 1 and the preceding year. The fact that these changes were not related to changes between year 1 and year 2 seemed even more reason to question whether they represented the impact of significant performance-related factors. Exactly what kind of factor(s) could lead to consistent effects within subjects, but inconsistent effects between subjects and between years in the same subject?

On balance, Cresswell concluded that the profile of changes across subjects was not defensible. While a degree of change is, no doubt, to be expected from one year to the next, the scale of changes proposed by the scrutineers across the different subjects seemed unreasonable. A more likely explanation was that large changes resulted from inappropriate grading decisions. Cresswell subsequently evaluated the scrutineers'

decisions under the assumption that ability profiles were generally similar from year 1 to year 2 across subjects. If this were true then differences in mean marks between years would be attributable to the difficulty of the examination papers alone (once differences in mark totals between years had been accounted for). As such, grade boundaries would have to be raised/lowered to maintain standards. Comparing the statistically predicted grade boundary movements (under the null hypothesis of no change in quality of the examination cohort) with the scrutineers' recommended movements, it appeared that the latter tended to be around 0.4 of the size of the former. It is important to stress that, in 77% of cases, the scrutineers did move the boundaries in the 'correct' direction. The implication seems to be that the subject matter experts were generally able to determine the *direction* in which grade boundaries needed to be moved to account for differences in examination difficulty between years, but not the *extent* of movement required. This effect is in keeping with earlier research by Good and Cresswell (1988) which showed that awarders' judgements (of the same grade boundaries on different tiers) tended to be lenient for lower tier papers and severe for higher tier papers.

Although there is not a great deal of research in this area, Cresswell's study is detailed and convincing. Moreover, its pessimistic evaluation of script scrutiny judgements ties in with research that has been conducted in other judgmental contexts (e.g. Berk, 1986). Importantly, Berk noted that, when guided by relevant statistical evidence, scrutineers' judgements of question difficulty could be made more consistent. Cresswell (1997), likewise, found this to be case with script scrutiny panels' judgements.

There are a number of possible explanations for the weakness of unguided judgmental recommendations. Some of these relate to issues discussed in relation to the Angoff procedure: for example, subject matter experts are not necessarily very good at appreciating the difficulty of individual questions. Other explanations are more specific to the script scrutiny format. Cresswell (1997) was able to identify many aspects of script scrutiny meetings that could be challenged on the basis of research on the social psychology of decision making in small groups. For example: the pressures to conform; the way in which evidence from available individual cases can override more general considerations; the tendency for groups to take riskier decisions that individuals acting alone; etc.. Berk (1986) was similarly concerned about the inappropriate impact of social psychological factors and noted that, while further research was needed, those responsible

for designing standard setting procedures should "devote serious attention" to recommendations for reducing undesirable effects.

Other inappropriate influences upon awarding decisions have been identified; in particular, the effect of consistency of performance within scripts (where an 'inconsistent script' is one with particularly high marks on certain questions and particularly low marks on others, in relation to all other scripts with the same total mark). Scharaschkin and Baird (forthcoming) found that A-level Biology awarders were more likely to classify inconsistent scripts as being worthy of lower grades than particularly consistent scripts or scripts of average consistency - despite the fact that all scripts had the same mark total. Likewise, A-level Sociology awarders were more likely to classify particularly consistent scripts as being worthy of higher grades than scripts of average consistency. It is not clear precisely what practical recommendations ought to follow from such findings. Perhaps the best advice would be that sample scripts were selected at random.

1.6 Statistical analysis of live data

1.6.1 Methodology

The statistical analysis of live data constitutes the second phase of the Level Confirmation Exercise. It is based on the assumption that final decisions concerning level boundaries should not be taken in the absence of detailed information concerning how well pupils have actually performed in the live tests. As such, it is not a method for maintaining standards, per se, but a method for collecting salient evidence.

The sampling technique in 1999 required each marker to return the marks of a randomly selected sample of (up to 22) pupils from one of the schools that she was marking. These data were analysed in order to determine, for example: extrapolated proportions of pupils at each level of each test given the draft level boundaries; gender breakdowns of performance; etc.. The statistical extrapolations also took into account the effects of borderline remarking on the final level distributions: the inevitable increase in 'pass-rates' resulting from the re-marking of candidates just below level thresholds.[14]

[14] Note that the evidence that was collected related purely to the main tests and not to the extension papers.

1.6.2 Conceptual rationale

Once again, the collection of live data is simply a way of obtaining relevant evidence rather than a technique for maintaining test standards. However, it makes a lot of sense to have an indication of the likely final level distributions within a test before level boundaries are agreed. If large, unexpected, discrepancies occur these can be fed into the final decision making meetings.

1.6.3 Technical issues

Research into the accuracy of the live data collection exercise in 1999 (Quinlan and Scharaschkin, 1999) demonstrated that, at both key stages, predictions were within 1% of the National Data Collection figures. This indicates that the procedures used for sampling were effective and that the data were robust for their intended usage.

On the other hand, it is interesting to note the surprise of the Rose report that live data entered into the discussion at all:

> "Papers circulated at the meeting gave detailed information on the effects of each threshold mark on the percentage of 11 year olds attaining the different levels, but this was not considered until after the main discussion. To an outsider, it was surprising that this 'reality check', as it was termed, should have entered into the debate. The same could be said of brief discussions on whether a particular outcome seemed plausible in relation to previous years' results. However, in neither English nor mathematics could this be considered to have contributed even marginally to the level thresholds set." (Rose, 1999, para.8.4)

It seems likely that the surprise of the Rose report is rooted in the common misconception that live data should not be used to set standards because they are irrelevant in a criterion referenced system. This belief is misplaced for two reasons. Firstly, a major theme of the present report is that national curriculum tests are not criterion referenced, but criterion related, and standards cannot be set purely on the basis of observed performances in scripts. Secondly, even if national curriculum tests were criterion referenced it would still be unlikely that *large* percentage changes in performance would genuinely occur from one year to the next, particularly if there were no clear patterns to them (e.g., between years,

subjects, papers or boundaries) or if no coherent explanations could be found for them. As such, there are very good reasons to analyse live data.

1.7 Final Level Threshold Setting

1.7.1 Methodology

The purpose of the Final Level Threshold Setting Meeting is to integrate all of the possible sources of evidence concerning which boundary decisions will ensure that standards have been maintained.[15] Once more, then, it is not actually a discrete technique for maintaining standards. However, it is certainly the most important stage of the standard maintenance process, as the validity of levels assigned to candidates is determined, ultimately, by the appropriateness of decisions made here. Indeed, it is arguably the most important stage of the test development process; as Morrison *et al.* (1995, p.180) note: "The reliability of the test is academic if its cutscore is set unreliably."

The decision-making panel for each meeting is convened by the QCA and composed of representatives from bodies involved in the test development process, principally: QCA staff, test development agency staff and the Lead Chief Marker. Independent observers are also invited to attend the meetings, for example, representatives from teacher associations, academic institutions, etc..

For each level boundary decision to be reached, the committee revisits all of the available evidence:

- recommended cut-scores from the Pre-test;
- recommended cut-scores from the Angoff procedure;
- the Draft Level Thresholds;
- recommended cut-scores from the Script Scrutiny;

---

[15] In actual fact, there is a prior meeting - the Draft Level Threshold Setting Meeting - which has responsibility for initial recommendations. This takes place a few months before the final meeting and is necessary because markers need to have some idea of the likely borderlines in order to begin borderline re-marking. The members of the committee are from the QCA and the test development agency and reach decisions based on information from statistical scaling and Angoff (using a report on these matters prepared by the test development agency).

- implications of different cut-scores from the Live Data Collection.

Only evidence from the script scrutiny and the live data collection is actually new; if these do not suggest problems with the Draft Level Thresholds then they can be accepted without further ado. If the new evidence suggests changes to the Draft Level Thresholds then consensus through discussion is used to determine the final boundary decisions.

1.7.2 Conceptual rationale

A conceptual rationale for this procedure might be taken from the sociological perspective upon the maintenance of examination standards (Cresswell, 1996; see also Wiliam, 1996). This approach developed in the wake of decades of research aimed at providing valid solutions to the problems of comparability. Many, if not most, researchers involved in such projects came to the conclusion that there are no techniques, procedures or methods that can necessarily warrant the conclusion that standards have been maintained. In short, they concluded that it is impossible to *guarantee* standards in any technical sense.

> "A failure to resolve such issues has produced a great deal of comparability research leading nowhere (e.g. Nuttall, 1979, 1986; Wood, 1976; Christie and Forrest, 1980; Newton, 1996). Earnest attempts to apply a wide variety of methodologies to the problem have tended to lead us back to the realization that no absolute conclusions can be drawn about comparability unless there is complete agreement about the values that are to be applied." Murphy *et al.* (1996, p. 282)

Cresswell and Wiliam (working independently) took the further step of proposing that the only way in which test standards can genuinely be considered to have been maintained over time is if a sociological stance is taken. Standards cannot be maintained in any technical sense without making untestable assumptions and value judgements. Therefore, if there are to be any acceptable pronouncements concerning standards, then there must be individuals or groups who have been empowered to reflect upon these untestable assumptions and to make their considered value judgements. Consequently - to the extent that standards can be said to be maintained at all - this is only because society is prepared to trust these individuals or groups to do their appointed tasks effectively and to the best of their abilities.

The process of determining whether test standards have been maintained is akin to the process of determining whether a person is guilty of murder or manslaughter (Newton, 1997a). For final decisions to be reached - which they must, and in a limited time frame - untestable assumptions must be made on the basis of inadequate evidence. In the courtroom the notion of 'due process' has been developed in order to encapsulate the way in which trials must be conducted to be deemed socially acceptable. The verdict of a jury might be considered analogous to the decisions reached by a panel empowered to make comparability judgements. If the conclusions of a 'comparability jury' are to be accepted by society then they must be seen to have been grounded in an appropriate 'due process' (Cizek, 1993; Cresswell, 1996; Whetton, 2000; Whetton *et al.*, 2000).

In many ways, then, the Final Level Threshold Setting Meeting is to the pronouncement 'test standards have been maintained' as the courtroom trial is to the verdict 'guilty'. The panel is empowered to declare that standards have been maintained in a particular test in the same way that a jury is empowered to declare that an individual is guilty. To the extent that society trusts this scenario, the declaration, in itself, might be said to create the 'social fact' of comparability/guilt (Wiliam, 1996).

Of course, this is a very general conceptual framework that does not at all specify the practical details. There need to be prior decisions concerning the nature of the 'due process', for example:

- from which groups of society should the 'comparability jury' be selected (e.g., test experts, subject specialists, test users, members of the public, pupils, combinations of such groups, etc.)?
- what evidence should be presented to members of the 'comparability jury' (e.g., complex statistical information, nothing more than recommended cut-scores from different methods, etc.)?
- should the 'comparability jury' be required to scrutinise the evidence in a particular manner (i.e., should they be required to scrutinise evidence systematically and, if so, what formal procedures should be involved)?

- should the 'comparability jury' be guided in their decisions by a 'comparability judge' (i.e., should there be a neutral expert to explain how different forms of evidence should be understood)?

- should there even be a 'comparability jury' (rather than a single 'comparability judge')?

- should it be possible to revisit/revise the decisions made by a 'comparability jury' (and, if so, under what conditions and who should be allowed to revisit/revise them)?

Most importantly, answers to these questions need to be seen to be *defensible* (see also Quinlan and Scharaschkin, 1999). Otherwise society will simply not invest the trust that is the fundamental requirement for the system to work. Current procedures for setting and maintaining test/exam standards sometimes have the appearance of being *ad hoc* and under-specified (at all levels of the UK assessment system, but particularly in national curriculum assessment). Such an appearance brings into question the defensibility of standards set.[16]

1.7.3 Technical issues

> "In [the Key Stage 2] English [Final Level Threshold Setting Meeting of 1999], debate centred on the reading test, where year-on-year differences are more marked than in writing, which has been judged on the same criteria since the tests were introduced. Provisional thresholds for writing were confirmed at all levels with little debate. In reading, however, the choice of texts and the questions set inevitably create more uncertainty over standards. At Level 3, for example, there was a difference of five marks between the threshold suggested by statistical comparisons with previous years and the judgements reached by teachers and senior markers. The markers' view prevailed, but not before reservations had been expressed by those responsible for the pre-testing procedures." (Rose, 1999, para.8.5)

[16] Of course, this is to side-step a prior (or at least a concurrent) problem, which is to clarify precisely what we mean by 'defensible' in the context of standard setting and maintenance.

In contrast to the recognised methods for determining cut-scores, which have each been the subject of continued scrutiny for many years, the *integration of evidence* from each of the methods (to determine final boundary decisions) has largely been overlooked.

Procedures are most tightly defined for GCSE and A-level grade awarding. The awarding committee is required to determine upper and lower boundary limits, through script scrutiny, and then to reach final recommendations using both professional judgement and statistical evidence. Statistical evidence is defined by the Code of Practice for awarding as "previous year's statistical outcomes of the component at the boundary in question, and information about changes in entry patterns, estimated grades and other technical data." (QCA, 1999, p.24). In actual fact, the Code specifies that the final grade boundary decision must be made by the awarding body's 'accountable officer' who evaluates the awarding committee's final recommendations "by reviewing all the evidence." (QCA, 1999, p.25). The point to notice is that, while the Code of Practice acknowledges the significance of the final decision, and of the multiple sources of evidence upon which that decision should be based, it gives no indication of the kinds of considerations and criteria that need to be reflected upon when making such decisions.

It may be that this is precisely as it should be: each decision should be taken on its individual merit and no guidelines laid down because they would only lead to inflexibility. Whether or not guidelines should be formalised, a debate concerning 'best practice' would seem appropriate if awarding decisions are to be seen to be defensible. There appear to be two major issues that would benefit from closer attention, especially in the national curriculum context. Firstly, with whom should responsibility for final decisions rest? Secondly, what practical advice can be given to facilitate the decision-making process?

In relation to the issue of responsibility there are actually wide variety of views to be found; we will consider the following:

1. Final decisions should be made by those who best understand the technicalities of linkage;

2. Final decisions should be made by those who best understand what is being linked;

3. Final decisions should be made by those who have ultimate responsibility for the assessment system;

4. Final decisions should be made by representatives from different parts of the system, collectively;

Argument 1 stresses that the task of linking tests is a highly complex one which requires a fundamental appreciation of the significance of assumptions that are made by different models. In the USA, where final decisions typically result from the use of a single agreed-upon methodological approach, this criterion probably wouldn't be of significance - participants don't need to understand the model being used, they simply need to follow accepted procedures appropriately. In the UK, though, emphasis is upon the integration of evidence from multiple sources; moreover, the task is specified as 'maintaining' rather than 'setting' standards, which is a subtle distinction, but one that both adds another level of complexity and that enhances the requirement for defensibility. There is a strong case to be made, then, that those responsible for making final decisions should be those who best understand the technicalities involved. On the other hand, this might result in statistically minded measurement professionals typically ending up with overall responsibility. Might they not be biased in favour of statistical evidence? Whether such an approach would lack face validity is an important consideration that needs to be addressed.

Argument 2 counters this with the proposal that final decisions should be taken by those who best understand what is being linked, i.e., the professionals who teach and informally assess pupils. This is what happens when the Angoff procedure, or the script scrutiny, is used as the sole method for setting cut-scores. The underlying model is strongly supported by Wiliam's construct referencing model, in which the shared 'inter-subjective' constructs of a community of professionals are deemed the most appropriate basis for defining standards (Wiliam, 1996). This relates to Cresswell's discussion of the 'connoisseurship' of examiners (Cresswell, 1994). However, as Massey notes:

> "... 'connoisseurship' is unlikely to suffice for teachers seeking the rationale for KS3 'standards'. ... Connoisseurship (or expert judgement) is likely to have a role in this, but the transparency of the national assessment system may require us to

describe how judgements help to establish the link between level descriptions and cut-scores: we may need to clarify and defend our procedures." (Massey, 1995, p.191)

If accountability for final decisions is to lie solely with the professional judgement of teachers, we need to be confident that such decisions can be shown to be technically adequate. Given the discussions of the preceding sections, it is not clear the extent to which such a defence could be made.

Argument 3 is similar to the model for GCSE awarding. There is a single point of accountability - the accountable officer - who represents the examining body and who has overall responsibility for standards within that agency. This is essentially a political decision, as there seems to be no requirement that the 'accountable officer' in the GCSE board should have any direct experience either of the technicalities of comparability or of teaching. This approach is attractive in the sense of not letting either the measurement professional or the teacher have final say. However, it is unattractive in the sense that an accountable officer with no detailed understanding of teaching or comparability must surely rely on others for advice when statistical and judgmental recommendations conflict.[17] Whether the accountable officer tends to favour advice from measurement professionals, or from teachers, or someone else entirely, will surely introduce an element of systematic bias into the process (whether justified or unjustified, conscious or unconscious).

Argument 4 is, in some senses, the democratic ideal. Ensure that everyone who has an opinion is allowed to air it and require that decisions be made through group consensus. This, of course, assumes that consensus can be reached. Bearing in mind the previous quotations from the Rose report, it does not appear that this always happens. Indeed, the question arises as to whether different parties are ever likely to reach genuine agreement when evidence from different methods conflicts. If not, then the final decision must either be made by a 'chairperson-judge' (in which case Argument 4 may end up amounting to either Argument 1, 2 or 3) or through a voting system. The vote might

[17] Unless, for example, the accountable officer were simply to apply a rule such as 'choose the recommendation that would lead to the smallest year-on-year change in proportion of pupils at each level, given the live data collected'.

involve all participants recommending their own boundary mark and the final decision could represent the mean, mode or median of these recommendations. Exactly who is invited to take part in the decision-making committee is, of course, central to the validity of Argument 4; this is particularly salient if a formal voting system is to be employed.

Adopting the courtroom analogy, we might argue that committee members for awarding Key Stage 2 national curriculum tests should be randomly selected from the general public and be presented with evidence from each of the various methods. Whether this would be seen as a sufficiently credible approach to maintaining standards is, essentially, an empirical issue. It might be. At the other extreme, committees might be constituted by experts, for example, one expert for each of the three methodologies used: scaling, Angoff and script scrutiny. Somewhere in between would be a committee comprised of 'knowledgeable professionals', for example: experienced Key Stage 2 teachers; testing agency representatives; QCA representatives; DfEE representatives; academics; etc..

Whichever form a committee took, it would seem sensible to ensure that all members were sufficiently knowledgeable about the general principles and practices underlying the different standard setting methods. In addition, it would be sensible to ensure that they were presented with sufficient evidence concerning the actual contexts in which the different methods' recommendations were generated (noting, for example, when assumptions had clearly been violated or when accepted procedures had not been followed). It would seem reasonable that this information be in documented form and presented to committee members in advance of the Final Level Threshold Setting Meeting. Such considerations would seem appropriate for each of the Arguments presented above.

While the issue of locating responsibility for final level boundaries is problematic, it is somewhat easier to solve than the issue of how those responsible should make their decisions. Interestingly, the courtroom analogy would suggest that this be left entirely to the accountable individual or committee, without scrutiny or requirement to defend any judgement. To the extent that this is acceptable in law it might also be deemed acceptable in standard setting contexts.

If this absolute under-specification was not deemed defensible, one way of circumscribing more tightly the Final Level Threshold Setting Meetings might be for formal weightings (in the statistical sense) to be assigned to the different types of evidence. The terms of reference for a committee might be to reach consensus concerning the relative weighting to be applied to the cut-score recommendations from each of the three methods. These would mathematically determine the final cut-score. If no consensus could be reached, then voting might be used (either by voting on different weighting models or by some kind of averaging of recommended weights across representatives). This circumscription would be intended to enhance the apparent 'due process' of the meeting and, hence, its defensibility. This type of approach should be discussed and researched more fully. However, it is possible that such an approach might have one major limitation. It is perfectly possible that a committee might want to recommend a final cut-score that could not be arrived at through any weighting of recommendations. For example, the Angoff, script scrutiny and scaling methods might all recommended a cut-score of 25. In contrast, evidence from the live data collection exercise, in light of the known methodological weaknesses of each of the three methods, might suggest that a more realistic cut-score would be 23. No weighting of the three recommendations could result in this cut-score. While the task of weighting evidence is quintessentially the purpose of the Final Level Threshold Setting Meeting, it may not always be appropriate to employ formal mathematical procedures. Whether the process can be adapted for more informal procedures remains a useful avenue for exploration.

Ignoring the problem of what procedures should be followed to arrive at cut-scores, there remains the underlying problem of what principles should guide decisions made between conflicting sources of evidence. When reflecting upon the dependability of recommendations from each of the methods, one important consideration may be the extent to which they are independent; although at first glance they appear to be, this is not necessarily the case. For example, the iterative Angoff procedure uses pre-test data to enhance the consistency of teachers' probability judgements. Note that, if this pre-test data were somehow suspect, then both the scaling and the Angoff recommendations might be affected. Furthermore, even the script scrutiny is not independent of the pre-test because the scripts are selected from a range that centres on the boundaries recommended by scaling (Quinlan and Scharaschkin, 1999). Finally, assuming the test potential model,

the logic of all three methods will be compromised when curricula (and even question formats) change significantly.

The most important considerations in assigning weight to different boundary recommendations will concern the manners in which the methods differ. In particular, the different assumptions that the models make and their relative plausibilities. Many of these issues have been discussed in previous sections of the present report. A wider debate, focusing on issues such as these, will be important in moving towards a more defensible system.

1.8 General conclusions of Section 1

The main conclusions to be drawn from the foregoing review are summarised below:

1. When a curriculum remains unchanged from one year to the next then there is a coherent theoretical framework through which the maintenance of test standards can be understood; the present report has called this 'test potential referencing'.

2. Test potential referencing supports a number of methods for generating cut-scores, in particular, the three approaches adopted at Key Stage 2 for English and science: statistical scaling; the iterative Angoff procedure; and script scrutiny.

3. All of the three approaches used at Key Stage 2 are limited because they each necessitate a variety of untestable assumptions; it is generally accepted, within the educational measurement community, that untestable assumptions are unavoidable in standard setting contexts.

4. As a consequence of untestable assumptions, there is no simple resolution when different methods recommend different cut-scores; final recommendations can only be made through value judgements concerning the relative plausibility of assumptions made by each method.

5. Further debate is needed to determine more precisely the procedures and criteria for dealing with conflicting recommendations; this is especially important in light of the conceptual and pragmatic limitations of each method that have been identified within the present report.

6. It is fair to argue that "there is no gold standard" (Brennan, 1998, p.9); however, that does not mean that standards are inherently arbitrary and indefensible, *as long as procedures are seen to be rationally derived, consistently applied and explicitly described* (Cizek, 1993).

7. Changes in curriculum undermine the coherency of test potential referencing as a conceptual framework through which the maintenance of test standards can be understood; the larger the curriculum change, the more indefensible any claim to have maintained test standards becomes.

2: Possible Revised or New Approaches

Two different responses to the present report will be outlined; both of these accept that, while the present system is generally defensible, improvements can be made. The first proposes that steps might be taken to make procedures more defensible within the general test potential referencing framework. The second proposes that serious advantages (and very few disadvantages) would accrue if national curriculum tests moved to a cohort referencing framework, indeed, a strong cohort referencing framework in which results were simply recorded as standard scores and levels were not awarded at all.

2.1 Make the system more defensible

There appear to be four ways in which current procedures for maintaining national curriculum test standards could be made more defensible:

1. Implement changes that we are confident will improve matters.
2. Investigate issues that remain unresolved.
3. Ensure consistent application of best practice (hence, defensibility) through explicit documentation.
4. Ensure procedures for maintaining standards are widely understood and generally accepted.

To some extent, these stages are progressive, for example, it might not be appropriate to codify best practice until further research had elucidated some fundamental issues. In other ways, the stages are iterative, for example, feedback on the general acceptability of procedures would need to be a constant feature of any development work.

2.1.1 Changes to present procedures that are worth debating

Consideration should be given to the following suggestions, each of which is intended to make the current procedures more effective.

Pre-test and scaling

a) Do not change items or test format after final pre-testing. Untestable assumptions are unavoidable whenever changes are made. *This is an obvious weakness that should be avoided if at all possible.* Likewise, marking schemes should not change. If they do, then pre-test scripts should be re-marked according to the new scheme and scaling re-computed on these marks. (A possible alternative, that would accommodate changes to a second pre-test, might be to scale cut-scores on the basis of a third pre-test.)

b) Take steps to minimise the 'pre-test effect'. One approach might be to ensure that a random sample of pupils took the pre-test instead of the live test, so the pre-test would effectively become the live test for those pupils (Whetton, 2000). Final levels for these pupils would be awarded on the basis of cut-scores scaled directly from the actual live test; in effect, these pupils would simply be taking the year 2 test, with year 2 cut-scores, in year 1. This would have to be considered carefully, though, as there might be legal problems, problems of persuading schools or problems of incorporating in statistical accountability tables.

   Another approach might be to administer two non-live tests at the pre-test stage. Kiek (1999) suggested requiring a sample of pupils, in year 2, to sit both the past paper from year 1 and the pre-test for year 3; instead of scaling cut-scores from year 2 to year 3 (as would traditionally happen) cut-scores would be scaled from year 1 to year 3. Again, such a proposal would have limitations. Firstly, it would mean that, each year, the sample of pupils would have to sit two non-live tests in addition to the live test (although a compromise might involve two samples, each taking half of the two non-live tests). More importantly, the year 1 test would already have entered the public domain, which is likely to compromise any scaling. An alternative would be to give year 3 and year 4 pre-tests to the sample of pupils in year 2. The year 3 cut-scores would have already been derived from a similar exercise in the previous year and these could then be scaled onto the year 4 pre-test. This seems to be the most practical solution, but it does have the disadvantage that tests would need to be fully developed, and cut-scores set, two full years in advance of the actual live

administration.[18] Indeed, questions would still remain concerning the appropriateness of scaling on the basis of non-live tests; equivalence based solely upon pre-test versions need not necessarily imply equivalence on live test versions.

c) Re-consider the value of the anchor test for Reading. Its size (hence reliability) and its inability to accommodate curriculum change mean that it is not an ideal tool for maintaining standards. An alternative might be to consider using one of the pre-tests that are rejected each year as an anchor. A rejected pre-test would be more appropriate because it: would measure the appropriate construct; would already be equated; and would be of full length. When significant curriculum change occurred, the most recent rejected pre-test could be adopted as the new anchor.

Angoff procedure

a) Give participants clearer guidance concerning the 'minimally competent pupil'. This could be achieved in the same way as for script scrutiny, i.e., through scripts at the appropriate boundary in the previous examination.

Script Scrutiny

a) Take steps to minimise inappropriate social psychological effects. For example: by not allowing formal or informal discussions of 'standards'(that are technically independent of the exercise) before the scrutiny begins; by encouraging formal written records of individual judgements; etc..

b) Consider requiring scrutineers to take account of mark distributions from live data collection (as is done in GCSE awarding). This would be likely to ensure more consistency of judgement from year to year. However, whether it would reduce the validity of the procedure should also be considered. If it was felt that validity would

[18] Kiek (1999) also recommended a number of models which rely on the use of a static anchor, or reference, test. Problems with this kind of approach have been discussed earlier in the present report. Both anchor and reference tests are particularly insensitive to curriculum change over time. Similarly, it must be assumed that the anchor, or reference, test measures the same construct as the real test. Kiek claimed that a correlation between real test and reference in the order of 0.6 is fairly good evidence of this; however, this level of correlation implies that about two-thirds of the variance in real test performance is *not* explained by performance in the reference. Surely a far higher correlation should be demonstrated before such an approach could be considered?

be compromised, then consideration should be given also to the Angoff procedure - perhaps pre-test data should not be used here for the same reason. Note that neither the Angoff nor the script scrutiny are used to generate final recommendations. As such, it might be deemed more valid to eliminate the integration of statistical information from both techniques.

c) Ensure that sample scripts are not unrepresentative; for example, that they are not all particularly inconsistent, or even particularly consistent, in terms of marks achieved across items.

Final Level Threshold Setting Meeting

a) Ensure that live data are presented to the meeting, and are presented at the beginning. The meeting should be guided in the importance of considering the potential impact of their decisions upon final distributions of levels.

b) The meeting should be convened primarily as a forum for *making* decisions and not for *ratifying* decisions made during the Draft Level Threshold Setting Meeting. As such, evidence from the scaling exercise, the Angoff procedure and the script scrutiny should all be evaluated in their own right and in light of the live data. (This is not to suggest that draft thresholds should not be determined at all, as these are necessary for borderline re-marking.)

c) The locus of accountability for maintaining standards should be made explicit and membership of decision-making panels should be reviewed. Particular attention should be paid to matching the panel membership to the form of decision making intended.

d) Establish 'best practice' guidelines for the meeting (which, if possible, will specify agreed procedures and principles for evaluation).

2.1.2 Issues that still need to be researched

There are still important issues that need to be researched. Without this research it will be hard to defend procedures for maintaining test standards in the national curriculum. The most significant issue to address is whether the judgmental methods have sufficient

validity, let alone reliability, to constitute meaningful sources of evidence. It could be argued 'the more evidence the better' regardless of quality; indeed, quality could explicitly be taken into account in a subsequent weighting process. On the other hand, it could be argued that paying any attention to essentially uninterpretable information - data that was merely masquerading as valid evidence - would be worse than not attending to it at all. As Murphy *et al.* comment:

> "A single method needs to have a reasonable credibility with respect to the questions which are being asked, in this case about comparability, for it to warrant inclusion even in a multi-method study." Murphy *et al.* (1996, p.288)

On the other hand, decisions as to what methods are used may also need to take into account the public understanding and perception of standards. Berk (1986) was quite insistent about this, as his two final 'criteria for defensibility' argue:

> "9. The method should be easy to interpret to lay people. Inasmuch as the *Standards* requires that the cutoff score along with the certification test results "be reported promptly to all appropriate parties, including students, parents, and teachers" (p.53), the method used to determine the cutoff should be interpretable and understandable to those audiences. Explanations of the method should be clear and conceptually simple for those professional educators and lay people who may need to defend it.
>
> 10. The method should be credible to lay people. A "statistically magical" method is typically not credible to lay people; neither is one that is conceptually confusing and intuitively unsound. A method that involves the input of representative samples of interested lay populations tends to possess greater credibility." (Berk, 1986, p.144)

If this is true, research is needed not simply into the technical adequacy of methods for maintaining standards, but also into the public perception and understanding of those standards. To determine technical adequacy, the present paper would recommend two particularly important areas for future research:

a) An investigation into the nature of 'pre-test effects'. It is important to be able to disentangle possible motivation effects from possible effects of practice or cramming

or curriculum coverage. Importantly, these effects may differ between subjects and key stages, which would mean a large-scale project. However, simply manipulating pre-testing procedures, such that half took the pre-test before the live and half after, might go some way to disentangling such effects.

b) Further investigation into the validity of decisions made during the Angoff exercise. We need reassurance that participants are capable of assessing 'absolute' difficulty, at least within acceptable limits.

2.1.3 Ensure consistent application of best practice through explicit documentation

There is a general need for better documentation within national curriculum assessment in order to facilitate best practice and to ensure defensibility. These needs are manifest in a number of ways, for example:

a) A need to document the general principles and practices appropriate for maintaining standards in national curriculum tests. This is, essentially, a call for a Code of Practice for national curriculum testing comparable to those introduced for GCSE and A/AS examinations in the mid-1990s.

b) A need to document specific details of the procedures actually followed. This might mean detailed reports upon all: pre-test development and scaling exercises; Angoff exercises; script scrutinies; live data collections; and final level threshold setting meetings. These need to be of the kind of detail that would enable replication.

Cizek referred to the importance of explicit documentation as an aspect of 'procedural validity'. He recommended that recording be as precise as possible, for instance:

> "In documenting a standard setting study, it is advisable that all aspects of the procedures used in actually implementing the chosen procedure be explicated in detail. The following list represents the minimum in terms of the kinds of information that should be included: the number and manner of selecting participants; the qualifications of participants; the qualifications of those designing and implementing the methodology; the materials used; the script or actual verbal instructions given to participants; key frameworks or conceptualizations

developed by participants [...]; the timeline, schedule of events, and actual agenda followed. [...] deviations from intended procedures should be noted and carefully explicated in the documentation of standard-setting study." (Cizek, 1996, p.16)

Indeed, some UK researchers have recommended an even higher level of explicit documentation:

> "Because British test developers are not required to demonstrate that their instruments meet high technical standards - Britain has no equivalent of the American Psychological Association's Standards for Educational and Psychological Testing (APA, 1985) - teachers and parents must accept, on trust, the quality of national tests." (Morrison and Wylie, 1999, pp. 92-3)

There is a clear need for better documentation of national curriculum standard setting principles, procedures and practices. However, it is important to note that this will require considerable time and effort and is likely to entail significant costs.

2.1.4 Promote the public understanding of assessment practices

The Rose report firmly recommended that more should be done to ensure that test users, and the public in general, gain a better understanding of assessment practices in the national curriculum:

> "It is important for everyone with an interest in the system to understand the processes that take place 'behind the scenes' to complete the test cycle, and why it is as it is. Relevant papers describing the test arrangements have been in the public domain for several years. The QCA also publishes a number of papers each year on the outcomes of the tests and the running of the system. Sadly, these are not as widely read as they might be. It is therefore not surprising that some of those critical of the tests think that what goes on is unduly shrouded in secrecy and therefore suspect." (Rose, 1999, para.4.6)

It is clear that if procedures are to be seen to be defensible they must not only be documented, but be seen to be documented. It must, at least, be generally known that relevant reports exist... even if they are rarely read. Much lip-service is given to importance of educating the public, but it seems unlikely that the public will be educated

through technical documents, however simply expressed. If the public is to be educated, this needs to be attempted realistically with the realisation that the public will only learn what they *want* to learn, *when* they want to learn it, and through the *format* that they prefer.

The price of education, however, is bare-faced honesty. Consider the following quotation from Brennan:

> "It is not much of an exaggeration to state that the assumptions in our [measurement] models are all false except for those that are true by definition. ... The crux of the matter, then, is not that we pick models with correct assumptions but rather that we recognize and acknowledge the fallibility of our models and assumptions, qualify results accordingly, and not mislead test users." (Brennan, 1998, p.5)

Dealing with honesty is a very difficult task. The principal problem is mis-representation. If the QCA were to publish a press release which accepted that all of its models were false, except for the ones that were true by definition, it would inevitably be taken by the press as an admission that standards had fallen (after all). Above all, it is probably this fact that explains the public perception that the examination boards and testing agencies are nefariously secretive. However, whether the boards and agencies could defend a relative lack of discussion of certain issues, on the basis that they might be mis-represented, is worth reflecting upon. Clearly, then, the problem is not simply to encourage education, but to discourage mis-education. The relationship between the national press and the agencies responsible for providing assessment information needs to be considered more closely in this respect.

2.2 Re-focus national curriculum assessment

> "Those who claim that one type of standard is better than another thoroughly miss the point - namely, that the type of standard should match the type of decision to be made." Brennan (1998, p.9)

The above quotation from Brennan cuts straight to the heart of the matter, by reminding us that the criteria for judging the defensibility of test standards are grounded in the types

of inferences and decisions that are made on the basis of results. The present report argues that national curriculum assessment would be very much easier to defend if it was re-focused. The point of re-focusing would be to ensure that test results were *sufficiently valid* and *sufficiently reliable* for *the uses to which they are put.* So why exactly do we need national curriculum assessment? The principal reasons fall under three main headings:

2.2.1 To indicate how well the education system is performing

Policy makers need to be sure that pupils are at least as well educated today as they have been in previous years, if not better educated. If educational standards are seen to fall then policy makers will be under pressure to intervene. In fact, many within the educational measurement community have concluded that test standards cannot be maintained over long periods of time and, as such, educational standards over time cannot be measured (e.g. Goldstein, 1979). Others have noted that, even if test standards could be maintained over time, this would still say little about the education system, per se, let alone about whether particular aspects of it were failing (e.g. Murphy, 1996; Newton, 1997a). Yet there still remains a political imperative to compare the present day effectiveness of the education system with that of yesteryear. Therefore, national curriculum assessment must approximate this as far as possible.

2.2.2 To indicate how well individual schools are performing

Since the early eighties there has been a statutory requirement upon secondary schools to publish examination results in aggregate form (GB. SI, 1981). This was the harbinger of a zeitgeist that was to develop during the latter years of the eighties and into the nineties: the idea that schools were in competition with each other. Indeed, the impetus to compete was even promoted amongst the primary sector with the Primary School Performance Tables (e.g. GB. DfEE, 1997). A reality of national curriculum assessment, therefore, is that results ought to be of sufficiently high reliability and validity to enable effective school comparison. Indeed, the assessment information generated needs to be sufficiently robust to enable comparison at the subject level (which carries the implication that standards between subjects ought to be equivalent to ensure that comparisons are seen to be fair). School comparison is an extremely 'high stakes' context and, as such, demands

extremely robust standards. Indeed, to the extent that individual teachers may be compared on the basis of pupil performance, national curriculum assessment needs to be even more rigorous, especially if teachers' salaries are somehow to be linked to test results (Rose, 1999, para.4.7).

2.2.3 To indicate how well individual pupils are performing

In educational terms, the most important function of national curriculum assessment is to provide feedback on individual pupils. This feedback is traditionally described as providing either formative or summative information. For summative purposes, it is often sufficient to know how well a pupil has done relative to other pupils. For instance, assessment information might be used to divide a class into high, middle and low sets; or to decide which pupils should have access to a local grammar school. For formative purposes, it is important to know more about the particular strengths and weaknesses of individual pupils in order that appropriate pedagogical intervention can be planned. In a sense, assessment for both summative and formative purposes is 'high stakes' - it is always important to draw appropriate inferences from test results - however, it is generally accepted that the stakes are higher when tests are used for selective purposes.

The point of the re-focusing proposed below is to draw distinctions between different assessment formats - linking each format to a specific intended assessment data usage - in order to meet the above demands more effectively. In the first instance, teacher assessment would be de-coupled from national tests. Instead of playing 'second fiddle' to the test results, teacher assessment would be the primary vehicle for reporting upon individual pupils. As well as satisfying the demand for formative information, it could also be used for 'low stakes' summative purposes. In fact, teacher assessment would remain essentially unchanged from its present form and would still report in terms of national curriculum levels. Teacher assessment would remain firmly criterion related.

From the point of view of the present report, the biggest change would concern tests results. The de-coupling from teacher assessment would be achieved when test results were cohort referenced. This would help to satisfy the demand for reliable and valid information upon which school comparison could be based. Results would not be reported in terms of national curriculum levels; instead, individual pupils would be

assigned a standard score which represented their performance in terms of their own year cohort.[19] Schools would be compared in terms of the average standard score of their pupils and performance in individual subjects could be compared likewise. While some might balk at the idea of reporting their standard scores back to individual pupils, there would actually be no need to feed test results back to pupils at all; moreover, there would be good reasons not to do so (see below). Notice that the only change to the present system is that there would be no need for the elaborate procedures for maintaining test standards from one year to the next (as no levels would be awarded). The tests would still assess the national curriculum in precisely the way they currently do.

Finally, a new national curriculum assessment format would be introduced to monitor the performance of the educational system over time. In practice, this would mean a return to *something like* the Assessment of Performance Unit (APU). In principle, though, the terms of reference would have to be carefully considered. The educational measurement community has come to view standards very differently over the past couple of decades; it is high time that the UK re-evaluated the monitoring of educational standards in this light. The impetus for a new Assessment of Performance Unit would be the perfect justification for instigating such a debate.

The rationale behind this proposed re-focusing of national curriculum assessment is to ensure that assessment information is fit for the purpose for which it is used. Fundamentally, the intention is to accommodate inevitable uncertainty by locating it in contexts that are most forgiving. Assigning pupils to discrete criterion related levels is an inherently ambiguous procedure, even if there are sound educational reasons for doing so. If we restrict the assignment of levels to teacher assessment - a relatively 'low stakes' situation in which few serious problems would arise through lack of comparability between years or even between institutions - then we minimise the impact of level mis-classification. Likewise, if we were to institute a new national curriculum assessment format for monitoring standards over time then we would be able to ensure that it had the characteristics best suited to the task. Inevitable uncertainty would still remain; but it

[19] Standard scores would be more useful for comparative purposes than percentile ranks, although the principle of cohort referencing would be the same. Raw scores in each subject might, for example, be converted to a distribution with a mean of 500 and a standard deviation of 50. The exact form of the scores would need to be agreed.

would be located in an assessment context that was specifically designed to be maximally forgiving of this uncertainty.

2.3 Further Discussion of Re-focussing

A fuller discussion of the pros and cons of the proposed re-focusing is presented below. Each of the three uses of assessment information is addressed in turn.

2.3.1 To indicate how well the education system is performing

The blanket testing of all pupils is not a good format for assessing improvements in the performance of cohorts over time. The task is complex enough in the best of situations, but in the compromised situation of national curriculum assessment standards over time are far from guaranteed. If standards are to be monitored in the traditional sense, then the most appropriate way to achieve this is: through 'low stakes' assessment; using samples of pupils rather than the entire population; by re-using items; and by ensuring that the entire curriculum is assessed each year using a range of test forms. The old Assessment of Performance Unit realised this as does the National Assessment of Educational Performance (NAEP) program in the States.

Having said this, even using the most appropriate methodology, there would still be problems in attempting to measure performance over time in the traditional sense. Such problems can be illustrated with respect to the Reading Tests carried out in Britain between the 1940s and 1970s (Start and Wells, 1972). Words such as 'mannequin parade' and 'wheelwright' clearly did not have the same significance in 1970 as in 1940 and therefore would not have had the same value in any test of reading ability (Goldstein, 1979; Nuttall, 1986).

This relates to the most significant of problems that must be faced by any attempt to monitor performance standards over time: the fact that curriculum change is not only inevitable, it is to expected. As society changes and (hence) curricula change, it is not simply the practical issue that tests have to change in order to accommodate this. The underlying problem is that the knowledge and skills valued and taught in a subject area at one point in time may be qualitatively different from those valued and taught in the same subject area a decade later. For example, pupils are now far less proficient in the use of

slide rules but far more proficient in the use of spreadsheets; does that mean that they are better, worse or no different in their ability to use calculating aids? We can make no such judgement, of course; technological developments mean that mathematics is simply done differently nowadays. We need a model for assessing change over time that can accommodate such complexities. It is not clear that this will necessarily, or at least exclusively, be the model used by the old APU or by NAEP. One thing seems clear, though: change over long periods of time cannot adequately be monitored using current national curriculum test results.

2.3.2 To indicate how well individual schools are performing

If an educational measurement expert was to design the ideal assessment format through which to realise cohort referencing then the result might look pretty similar to our national curriculum tests: there is a single curriculum for each subject which is (more or less) studied by everyone; all pupils in each year group are supposed to be assessed in all subjects; and test results are rarely, if ever, used to compare individual pupils from different cohorts. If it wasn't for the fact that strong criterion referencing was fundamental to the development of national curriculum assessment, we might have decided to cohort reference the tests from the outset![20]

Apart from an historic attachment to the discredited notion of strong criterion referencing, the problem with the proposal to cohort reference test results is that they could not be used to indicate *absolute* improvements in performance over time for individual institutions. Having said that, they would still indicate *relative* improvements in performance over time (relative, that is, to other schools) and this is the information that is of fundamental importance when comparing between institutions. Indeed, cohort referenced results would present this relative information in a form that was uncontaminated by variability in the maintenance of test standards from one year to the next. So, if a school's average Key Stage 2 standard score had increased from one year to

[20] Interestingly, commenting upon the inherent unreliability of methods for establishing comparability, Goldstein (1986) recommended that even 16+ examinations should be cohort referenced. Because of the different examination boards and curricula for the GCSE, and because of the fact that only a small proportion of the cohort opt for each examination subject (in a non-random manner), cohort referencing at this level *would* be problematic. However, Goldstein's response was that it would be more honest to adopt

the next, this would be direct evidence that its standing had improved in relation to other schools. Under the present system, if a school's proportion of level 4+ pupils had increased from one year to the next it would not be immediately obvious whether this was because the school had done particularly well, or because the cohort had improved generally, or both. In this sense, the information provided by cohort referenced results would be *more* appropriate for comparing performance between institutions. Indeed, it would be better in other senses: comparisons would be based upon a finer scale, standard scores rather than crude levels; standard errors of measurement could be routinely presented; and, as already mentioned, the scale would not be contaminated by variability in the maintenance of test standards.

On this matter of variability, Morrison and Wylie (1999) claimed that national curriculum levels have large regions of uncertainty associated with them, such that 'true' level boundaries might be as many as a few marks either side of the level boundaries actually set. In national curriculum terms, a few marks either side of a level boundary might mean a difference of, say, 15% in the proportion of Key Stage 2 pupils at level 4+. As Morrison and Wylie note, there are conceptual problems in speaking of cut-score 'error', as though we could ever speak of a 'true' cut-score (c.f. Dwyer, 1996). However, there is no doubt that level setting inevitably embraces variability. Indeed, it is accepted that certain standard setting procedures will tend to produce lower cut-scores than other, equally valid, ones (e.g. Jaeger, 1989). With this in mind, there are very good reasons to agree with those who argue that levels should not be set at all unless absolutely necessary (e.g. Glass, 1978). While national curriculum test results are necessary for accountability purposes, it is not at all clear that levels need to be attached to them. Indeed, the very fact that they are to be used for 'high stakes' accountability purposes would argue against the attachment of levels.

A further benefit of reporting standard scores is that the subjects assessed would be treated equally. What could be said when, in the first year of testing at a particular key stage, the proportion of pupils at a particular level was markedly higher in one subject than in another? Very little. Such standard setting decisions are essentially judgement

---

this approach, be open about its weaknesses, and then "the onus for a valid interpretation of the examination results would rest with the user rather than the present somewhat shaky comparability procedures" (p.183).

based and there are no alternative frameworks from which to argue that the initial recommendations are inappropriate. In principle, though, it is an odd way to begin. More importantly, it would mean that standards would have different meanings in different subjects (and would continue to have different meanings as long as standards were maintained). This appears somewhat inappropriate when results are used to hold subject teachers accountable. In the proposed cohort referenced framework, results would be directly comparable across subjects.[21]

Despite notable advantages, the problem remains that cohort referenced results would not provide information concerning *absolute* improvements in performance over time of individual institutions. While this need not be a problem for the setting of national performance targets (as these could be monitored by the new Assessment of Performance Unit), it does mean that target setting at the school and LEA level would need to be reconceptualised. There might be a number of ways in which this could be achieved. For instance, targets could be set purely on the basis of relative standards. Under the present system, improvement amounts to, for example, having to get an additional two Key Stage 2 pupils to the level 4 threshold. Under the re-focused system, improvement might mean increasing the average standard score of Key Stage 2 pupils by, for example, 10 points. Alternatively, individual school performance targets might still be set on the basis of levels, but using teacher assessment data. While this might suggest a tightening of moderation procedures, triangulation between a school's average standard score, the national monitoring exercise and teacher assessment would indicate whether the latter was going seriously awry. Of course, there is no reason why schools should not have performance targets expressed both in terms of average standard score and in terms of teacher assessment levels.

Judging the performance of institutions on the basis of average standard scores would also go a considerable way towards dissipating some of the negative consequences of expressing performance targets in terms of levels. The most important consequence is that teachers would not concentrate their attention on borderline pupils. There would be no borderline pupils - as far as the tests were concerned - and there would be no point in

---

[21] Although it is possible that different *distributions* of performance might still be observed in different subjects.

cramming them through the tests at the expense of other pupils. The cohort referenced approach would also mean that attention would not be focused on comparability to the detriment of other important issues: as comparability between years and subjects would no longer be an issue, attention could be re-focused elsewhere. Reducing the burden of pre-testing would mean that time and energy could be devoted to other (arguably) more important issues, such as the reliability of the national curriculum tests and their marking.

Finally, it is important to note that comparing schools in terms of average standard score would lead directly to transparent (if somewhat simplistic) value added analyses: the extent to which the average standard score of a class improved from Key Stage 1 to Key Stage 2 could be viewed as a direct index of the value added by a school. More complex and accurate measures based on multi-level modelling would also benefit from the increased variance in scores. (It is worth noting that all current value added analyses are essentially cohort referenced anyway. In any one year, the only issue of relevance to the statistical calculations is how well one school has performed in relation to all others in that year.)

2.3.3 To indicate how well individual pupils are performing

Perhaps the most significant failing of national curriculum assessment is the way in which teacher assessment has been undervalued (Stobart, 1999). Of course, if the reason why we have national tests at all is because we don't trust teachers to set standards consistently, then this failure is easy to understand. Exactly what messages are sent to teachers and parents when test and teacher assessment levels are in conflict? Teacher assessment needs to be re-evaluated and teachers need to be re-empowered. The proposed re-focusing would achieve this by ensuring that teacher assessment was the primary vehicle for reporting upon pupil achievement. It is often said that the strength of teacher assessment is reliability for individual pupils (owing to multiple assessment opportunities) while its weakness is consistency across teachers (owing to less than perfect moderation). Yet, when reporting individual levels to pupils and their parents, the potential impact of any teacher-level bias is relatively insignificant. This is an assessment context which is quite forgiving of the inevitable variability between teachers. This is not to say that moderation is not important; it is simply to say that reporting upon the

performance of individual pupils is a relatively 'low stakes' situation in which few harmful consequences of poor moderation would arise.

Where individual pupil achievement data was needed for 'high stakes' purposes, this would be available in the form of a national standard score. This information would be available to teachers, schools and LEAs despite not necessarily being made available to pupils or their parents. Teachers would therefore have access to information concerning the *relative* performance of their pupils which would complement their own *absolute* assessments.

A key advantage of not reporting test results to pupils might be to lower the testing stakes for them. This would be a constructive response to recent reports in the national press of 'test stress' amongst even the youngest of pupils (Stobart, 1999). In addition, not reporting results to pupils would also mean that they were not forced to think of themselves in terms of a de-contextualised score - no doubt a demoralising experience for anyone below the mean (see also Gipps, 1992). On the other hand, whether there would be a public demand for the release of pupils' standard scores is not clear. It might be that parents are as keen to know how well pupils are doing in relative terms as in absolute terms (indeed, relative information might be of more interest to parents). If standard scores were to be reported to pupils and parents this might risk de-motivating pupils who remained at the same position below the mean from one key stage to the next. Note that the supposedly motivating effect of reporting absolute achievement (which should increase over time for all pupils) was one of the main intentions behind the original Task Group for Assessment and Testing recommendation of a 10 level system. The potential effect of de-motivating pupils would need to be seriously considered before test results were released to parents or pupils. (It is important to recall, of course, that pupils and parents would still have access to teacher assessment data, which does describe absolute progress, regardless of whether standard scores were also released.)

A final comment must be made concerning the proposed re-focusing of national curriculum assessment. Cohort referencing national curriculum tests would mean that the validity of results would not be compromised in the least by curriculum change - even radical curriculum change. Thus, accountability judgements would be safeguarded from the most serious challenge to validity under the present system. Of course, curriculum

change would still have to be accommodated within the teacher assessment format and would have to be dealt with most directly by the new Assessment of Performance Unit.

Stobart (1999) outlined numerous very real threats to the validity of national curriculum assessment. The present recommendation is a direct response both to the general problem that he raised and to many of the specific ones. Stobart's conclusion was that the threats can best be minimised by ensuring that the two components, of teacher assessment and testing, "are kept 'in harness' - which may prove difficult given the current 'managerial' emphasis on test results" (p.12). The present proposal recommends precisely the opposite: the threats will only be overcome when the 'harnesses' are cut.

3: Policy Directions: Maintaining test standards and monitoring educational standards

> "Comparability can only be rough and ready, and is seldom as important as it is made out to be. More should be done to expose this situation, but one should be alive to the danger that if one rocks the boat too much there will be pressure for standardising syllabuses, and prescribing criteria that should be tested in any [...] examination." Nuttall (1979, p.58)

Over two decades have passed since Desmond Nuttall wrote these words. Comparability is, perhaps, higher on the agenda now than it has ever been; national curriculum assessment has both standardised syllabuses and prescribed criteria. Yet we are still no closer to solving the problems of comparability.

Yet we are closer to *understanding* the problems of comparability. In the intervening decade new perspectives have had a major impact on the way in which assessment experts view standard setting in both the USA (e.g. Cizek, 1993; Brennan, 1998) and in the UK (e.g. Cresswell, 1996; Wiliam, 1996). Gone are hopes that objective technical solutions might be found. Here to stay are claims that standards can only be set and maintained through value judgements, either (indirectly) concerning the plausibility of untestable assumptions or (directly) concerning the relative 'worth' of observed performances in the context of a changing society.

This is not a counsel of despair. It is simply a call to bring practice in line with theory. The essence of the new paradigm is defensibility. Value judgements underlying methods for maintaining test standards need be neither arbitrary nor capricious. So the most important task for the testing agency of the 21st century is to ensure that procedures are sufficiently rigorous to ensure that decisions are made with minimum caprice and maximum rationality. Standards must be set rationally, but they must also be seen to be set rationally. As the essence of the new paradigm is defensibility, it is important that the system as a whole, as well as individual decisions, are open to scrutiny. Above all else, this means that the full documentation of procedures used to maintain standards is crucial.

The fact that the 'old' (primarily statistical) techniques for maintaining standards are no longer believed to offer 'objective' solutions does not mean that they are necessarily

redundant. Nor does the more recent emphasis upon value judgements mean that the 'new' (primarily judgmental) techniques are necessarily more appropriate. The choice of techniques used to maintain standards must be driven by the uses to which results will be put. Where stakes are high, it may be that the largely known limitations of statistical techniques may be preferable to (and more defensible than) the largely unknown limitations of expert judgement; particularly if the untestable assumptions of the statistical techniques can be kept to a minimum. On the other hand, it may be easier to defend a limited approach that has face validity than one that does not. Thus, fallible subject matter experts may, perhaps, be seen as more credible than fallible statistical models. Whatever the outcome, the choice, or rapprochement, between subject matter experts and statistical models, needs to be approached in a manner that takes into account both technical fidelity and social acceptability.

The biggest challenge to the technical fidelity of procedures for maintaining test standards is curriculum change. This is true for both statistical and judgmental methods. As curricula change, in content or in emphasis, there comes a point when it is no longer meaningful to speak as if there were a coherent 'thing' to be maintained from one test to the next. At one extreme, the problem might be described as wanting to compare chalk with cheese (e.g. Wood, 1976). However, problems even arise at the other end of the extreme. While the curriculum for a certain subject area may appear to change only subtly from year to year, even these changes may be problematic. For example, imagine that a mathematics curriculum changed from one year to the next by replacing an 'algorithmic computation' section with a 'data handling' section (where both sections represented, say, 5 marks on the final maths test from a total of 100). If the students who tended to be good at algorithmic computation were not necessarily those who tended to be good at data handling then the year 2 test would be assessing a slightly different construct from the year 1 test. The problems of comparability turn from being technical to being conceptual: what do we mean by comparability when two tests measure different constructs? There are no obvious solutions to such problems. When we cannot even define what it means for two tests to be precisely comparable it is obvious that there can be no statistical nor judgmental techniques for guarantee test standards. The implication is that, when curriculum change is relatively small, the approximation of comparability

can be reasonably precise (at least in principle). However, when curriculum change is relatively large, the approximation of comparability is always open to question.

In practice, this tends to mean that test standards are generally robust enough to trust comparisons between one year and the next. However, the same cannot necessarily be said about comparisons between one decade and the next, as minor 'drifts' in standards (that would be negligible between years) can become significant over a longer period of time. Moreover, there are reasons to believe that systematic 'drifts' in standards may be a particular problem in relation to public tests and examinations. Perhaps the most important reason for this is the tendency - which is arguably quite appropriate - to give candidates the 'benefit of the doubt' in many circumstances.[22]

In short, there are serious challenges when test standards are to be maintained over time, both in practice (e.g., pre-test effects) and in theory (e.g., curriculum change). Indeed, we need to consider whether more is being demanded of comparability than can realistically be delivered. Ought we really aspire to monitor educational standards over time? If it really is imperative to attempt the task then we need to reflect upon whether the present arrangements are as robust as they possibly can be. The present report suggests that, if educational standards are to be monitored over *long* periods of time, *in relation to national performance targets*, then it would be more appropriate to adopt a methodology more akin to that of the American NAEP (bearing in mind lessons from the APU experience). The present report also suggests that, if educational standards are to be monitored over *short* periods of time, *in relation to individual school performance targets*, then it would be more appropriate to frame these in terms of relative rather than absolute comparisons.

---

[22] For example: choosing the lower of two cut-scores when percentile scaling suggests either of two possibilities or (at GCSE and A/AS) when different methods for combining cut-scores across components disagree; percentile scaling from a previous year's test results that have been inflated by borderline re-marking (which leads to up-grades but not down-grades); a tendency amongst scrutineers to choose the lower of two cut-scores when in doubt.

References

ANGOFF, W.H. (1971). 'Scales, norms, and equivalent scores.' In: THORNDIKE, R.L. (Ed) *Educational Measurement.* Second edn. Washington, DC: American Council on Education.

BAIRD, J., CRESSWELL, M.J. and NEWTON, P.E. (2000, forthcoming). 'Would the *real* gold standard please step forward?' *Research Papers in Education*, **15**, 2, 213-29.

BARDELL, G.S., FORREST, G.M. and SHOESMITH, D.J. (1978). *Comparability in GCE: a Review of the Boards' Studies, 1964-1977.* Manchester: Joint Matriculation Board.

BEJAR, I.I. (1983). 'Subject matter experts' assessment of item statistics', *Applied Psychological Measurement*, **7**, 3, 303-10.

BERK, R.A. (1986). 'A consumer's guide to setting performance standards on criterion-referenced tests', *Review of Educational Research*, **56**, 1, 137-72.

BRENNAN, R.L.(1998). 'Misconceptions at the intersection of measurement theory and practice', *Educational Measurement: Issues and Practice*, **17**, 1, 5-9, 30.

CHRISTIE, T. and FORREST, G.M. (1980). *Standards at GCE A-level: 1963 and 1973* (Schools Council Research Studies). London: Macmillan Education.

CHRISTIE, T. and FORREST, G.M. (1981). *Defining Public Examination Standards* (Schools Council Research Studies). London: Macmillan Education.

CIZEK, G.J. (1993). 'Reconsidering standards and criteria', *Journal of Educational Measurement*, **30**, 2, 93-106.

CIZEK, G.J. (1996). 'Standard-setting guidelines', *Educational Measurement: Issues and Practice*, **15**, 1, 13-21.

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*, 1, 1, 45-61.

CRESSWELL, M.J. (1996). 'Defining, setting and maintaining standards in curriculum-embedded examinations: judgmental and statistical approaches.' In: GOLDSTEIN, H. and LEWIS, T. (Eds) *Assessment: Problems, Developments and Statistical Issues*. Chichester. John Wiley & Sons.

CRESSWELL, M.J. (1997). '"Judge not, that ye be not judged." Some findings from the grading processes project'. Paper presented at the Associated Examining Board Research Seminar held at Regent's College, London, 21 November.

CRESSWELL, M.J. (2000). 'The role of public examinations in defining and monitoring standards.' In: GOLDSTEIN, H. and HEATH, A. (Eds) *Educational Standards* (Proceedings of the British Academy Vol.102). Oxford: Oxford University Press.

CRESSWELL, M.J. and HOUSTON, J.G. (1991). 'Assessment of the National Curriculum - some fundamental considerations', *Educational Review*, **43**, 1, 63-78.

DAVIS, A. (1998). *The Limits of Educational Assessment*. Oxford: Blackwell.

DEARING, R. (1993). *The National Curriculum and Its Assessment: Final Report*. London: SCAA.

DWYER, C.A. (1996). 'Cut scores and testing: statistics, judgment, truth, and error', *Psychological Assessment*, **8**, 4, 360-62.

FEUER, J., HOLLAND, P.W., BERTENTHAL, M.W., HEMPHILL, C. and GREEN, B.F. (1998). *Equivalency and Linkage of Educational Tests: Interim Report*. Washington, DC: National Academy Press.

FORREST, G.M. and SHOESMITH, D.J. (1985). *A Second Review of GCE Comparability Studies*. Manchester: Joint Matriculation Board.

FOWLES, D.E. (1995). *A Comparability Study in Advanced Level Physics: a Study Based on the Summer 1994 and 1990 Examinations*. Manchester: Northern Examinations and Assessment Board.

FOXMAN, D., RUDDOCK, G., JOFFE, L., MASON, K., MITCHELL, P. and SEXTON, B. (1985). *A Review of Monitoring in Mathematics 1978-1982: Part 2*. London: DES, Assessment of Performance Unit.

GIPPS, C.V. (1992). 'National Curriculum assessment: a research agenda', *British Educational Research Journal*, **18**, 3, 277-86.

GLASS, G.V. (1978). 'Standards and criteria', *Journal of Educational Measurement*, **15**, 4, 237-61.

GOLDSTEIN, H. (1979). 'Changing educational standards: a fruitless search', *Journal of the National Association of Inspectors and Educational Advisers*, 11, 18-19.

GOLDSTEIN, H. (1986). 'Models for equating test scores and studying the comparability of public examinations.' In: NUTTALL, D.L (Ed) *Assessing Educational Achievement.* London: Falmer Press.

GOLDSTEIN, H. (1996). 'Statistical and psychometric models for assessment.' In: GOLDSTEIN, H. and LEWIS, T. (Eds) *Assessment: Problems, Developments and Statistical Issues*. Chichester: John Wiley & Sons.

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

GREAT BRITAIN. DEPARTMENT FOR EDUCATION AND EMPLOYMENT (1997). *Primary School Performance Tables 1996 Key Stage 2 Results*. London. DfEE.

GREAT BRITAIN. STATUTORY INSTRUMENTS (1981). *Education (School Information) Regulations 1981* (SI 630/1981). London. HMSO.

GREEN, B.F. (1995). 'Comparability of scores from performance assessments', *Educational Measurement: Issues and Practice*, **14**, 4, 13-15, 24.

IMPARA, J.C. and PLAKE, B.S. (1998). 'Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method', *Journal of Educational Measurement*, **35**, 1, 69-81.

JAEGER, R.M. (1989). 'Certification of student competence.' In: LINN, R.L. (Ed) *Educational Measurement.* Third edn. New York, NY: MacMillan.

KIEK, L.A. (1999). *The Pre-test Effect: How It Affects Cut-score Setting, and How It Can Be Overcome Using Some Alternative 'Standards Fixing' Models.* Cambridge: University of Cambridge Local Examinations Syndicate.

KOLEN, M.J and BRENNAN, R.L. (1995). *Test Equating: Methods and Practices*. New York, NY: Springer Verlag.

MASSEY, A.J. (1995). 'Criterion-related test development and national test standards', *Assessment in Education*, **2**, 2, 187-203.

MORRISON, H., HEALY, J. and WYLIE, C. (1995). 'Teacher knows best: a solution to the marks-to-levels problem in National Curriculum testing', *British Educational Research Journal*, **21**, 2, 175-82.

MORRISON, H.G., BUSCH, J.C. and D'ARCY, J. (1994). 'Setting reliable National Curriculum standards: a guide to the Angoff procedure', *Assessment in Education*, **1**, 2, 181-99.

MORRISON, H.G. and WYLIE, E.C. (1999). 'Why National Curriculum testing is founded on a methodological thought disorder', *Evaluation and Research in Education*, **13**, 2, 92-105.

MURPHY, R. (1996). 'Like a bridge over troubled water: realising the potential of educational research', *British Educational Research Journal*, **22**, 1, 3-15.

MURPHY, R., WILMUT, J. and WOOD, R. (1996). 'Monitoring A level standards: tests, grades and other approximations', *The Curriculum Journal*, **7**, 3, 279-91.

NEWTON, P. (1997a). 'Examining standards over time', *Research Papers in Education*, **12**, 3, 227-48.

NEWTON, P.E. (1997b). 'Measuring comparability of standards between subjects: why our statistical techniques do not make the grade', *British Educational Research Journal*, **23**, 4, 433-49.

NUTTALL, D.L. (1986). 'Problems in the measurement of change.' In: NUTTALL, D.L. (Ed). *Assessing Educational Achievement*. London: Falmer Press.

POPHAM, W.J. (1981). *Modern Educational Measurement*. Englewood Cliffs, NJ: Prentice-Hall.

QUALIFICATIONS AND CURRICULUM AUTHORITY (1999). *GCSE and GCE A/AS Code of Practice*. London: QCA.

QUINLAN, M. (1995). *A Comparability Study in Advanced Level Physics: a Study Based on the Summer 1994 and 1989 Examinations*. London: University of London Examinations and Assessment Council.

QUINLAN, M. and SCHARASCHKIN, A. (1999). 'National Curriculum testing: problems and practicalities.' Paper presented at the British Educational Research Association Annual Conference, University of Sussex, Brighton, 2-5 September.

ROSE, J. (1999). *Weighing the Baby: the Report of the Independent Scrutiny Panel on the 1999 Key Stage 2 National Curriculum Tests in English and Mathematics*. London: DfEE.

RUDDOCK, G. and TOMLINS, B. (1993). *Evaluation of National Curriculum Assessment in Mathematics and Science at Key Stage 3: the 1992 National Pilot.* London: SEAC.

SAINSBURY, M. and SIZMUR, S. (1998). 'Level descriptions in the National Curriculum: what kind of criterion referencing is this?' *Oxford Review of Education*, **24**, 2, 181-93.

SCHAGEN, I. and HUTCHISON, D. (1994). 'Measuring the reliability of National Curriculum assessment', *Educational Research*, **36**, 3, 211-21.

SCHARASCHKIN, A. (1999). Comparability of Public Examination Standards: Some Theoretical and Practical Considerations. Unpublished paper.

SCHARASCHKIN, A. and BAIRD, J (forthcoming). 'The effects of consistency of performance on A level examiners' judgements of standards', *British Educational Research Journal.*

SCHOOL CURRICULUM AND ASSESSMENT AUTHORITY (1996). *Standards in Public Examinations 1975 to 1995: a Report on English, Mathematics and Chemistry Examinations Over Time.* London. SCAA.

SHEPARD, L. (1980). 'Standard setting issues and methods', *Applied Psychological Measurement*, **4**, 4, 447-67.

SHORROCKS-TAYLOR, D. (1999). *National Testing: Past, Present and Future* (Issues in Assessment and Testing). Leicester: BPS Books.

SIZMUR, S. (1997). 'Look back in Angoff: a cautionary tale', *British Educational Research Journal*, **23**, 1, 3-13.

START, K.B. AND WELLS, B.K. (1972). *The Trend of Reading Standards.* Slough: NFER.

STOBART, G. (1999). 'The validity of National Curriculum assessment.' Paper presented at the British Educational Research Association Annual Conference, University of Sussex, Brighton, 2-5 September.

WAINER, H. (1999). 'Comparing the incomparable: an essay on the importance of big assumptions and scant evidence', *Educational Measurement: Issues and Practice*, **18**, 4, 10-16.

WHETTON, C. (2000). 'Pre-testing in national curriculum tests.' Presentation to the Advisory Group on Research into Assessment and Qualifications. 1 March 2000. Qualifications and Curriculum Authority. London.

WHETTON, C., TWIST, L. and SAINSBURY, M. (2000, forthcoming). 'National tests and target setting: maintaining consistent standards.' Paper presented at the American Educational Research Association Annual Conference, New Orleans, 24-28 April.

WILIAM, D. (1996). 'Standards in examinations: a matter of trust?' *The Curriculum Journal*, 7, 3, 293-306.

WOOD, R. (1976). 'Your chemistry equals my French', *Times Educ. Suppl.*, 30 July.

Appendix 1: An illustration of the test development and awarding process[23]

The following acronyms have been used:

TDT - Test Development Team (e.g., NFER officers - research, statistics)
QCA - QCA officers (key stage, subject, statistical)
TRG - Test Review Group (e.g., QCA subject officers, teachers, academics)
EMA - External Marking Agency (e.g., SEG officers - key stage, subject; external markers)

Stage (time-scale)	**Procedures employed to develop and award the Key Stage 2 tests**	**Methods employed to ensure comparability with previous years**
Item writing (May-Sep 1998)	The development of a large pool of questions (e.g., 4 times the number of items that will eventually be needed) and the selection of texts where relevant.	The use of experts (TDT and QCA/TRG) who are familiar with the standards applied in previous years enables pilot items to be pitched at the appropriate level.
Pilot trials (Oct-Nov 1998)	A small pilot (e.g., 20 schools) is conducted to indicate the likely functioning of the potential test items. This leads to a reduction in size of the question pool and a preliminary honing of remaining questions.	The use of statistical information on the facility and discrimination of test items gives a quantitative indication of the extent to which questions are pitched appropriately.
First pre-test (Jan-Mar 1999)	Potential versions of test papers (e.g., 5 versions) are administered to a representative sample of pupils (e.g., 400 per version from year 6 and 7). Initial mark schemes are developed by the TDT, who may meet with senior markers (employed by EMA) to mark responses. Performance on the tests/items is analysed by TDT statisticians. The question pool is reduced in size further and agreement is reached upon one preferred and one back-up version of the pre-test. Following feedback from the TRG further honing of remaining questions may take place.	Further statistical information concerning the facility and discrimination of test items is collated. Test performance is scrutinised by TDT statisticians to determine whether similar mean marks are obtained on the current tests as were obtained in previous years. The TDT also compares test scores with teacher assessments of pupils and investigates teachers' views of the tests.

[23] Particular subjects may vary slightly from this example.

Second pre-test (Apr-Jul 1999)	The final pre-test versions (sometimes 2 versions of each test) are administered to representative samples of year 6 pupils (e.g., 1,500) a few weeks after they sit their actual Key Stage 2 tests. This results in a final selection of test versions which will go live. Senior markers further hone mark schemes and final (minor) amendments to test questions are made. Marker training materials are prepared by the TDT's Lead Chief Marker, the TDT, QCA and representatives from the EMA.	Statistical scaling, by TDT statisticians, enables marks obtained by pupils on the pre-test to be compared directly with marks/levels achieved by the same pupils in their end of Key Stage 2 tests. This results in recommended level boundaries for the pre-test that are based directly upon evidence of performance on the previous year's test. (Equipercentile scaling methods are used to establish comparability.)
Angoff scaling procedure (late 1999)	For each subject, the TDT gives a sample of teachers the final pre-test version and asks them to consider each question individually. They are required to judge the probability of success on each question that can be expected of a child performing at the borderline of each level. The teachers' responses are aggregated (across items and then teachers) to generate a prediction for the test performance of a typical borderline pupil. This is repeated for all level boundaries.	The use of a sample of subject specialists - teachers not involved in the test development process - enables level boundary recommendations to be determined on the basis of the item-level judgement of experts. (The process assumes that subject experts possess representations of 'borderline' pupils from which they are able to extrapolate hypothetical performances.) It requires decisions to be made by individuals without necessarily reaching any consensus.
Draft Level Threshold Setting (early 2000)	A report is prepared by the TDT (for each final pre-test version) that collates all judgmental and statistical evidence concerning its standard. The report is reviewed by the QCA and the TDT, focusing on any potentially conflicting recommendations and how best to resolve them.	The group consensus - of experts involved in the test development process - is used to integrate various sources of information concerning the potential location of level boundaries.
Level confirmation exercise - **Script Scrutiny** (June 2000)	After test administration and marking, EMAs convene Script Scrutiny exercises in which Lead Chief Markers chair meetings of Chief and Deputy Chief Markers to reach independent recommendations for the level boundary locations. Markers first scrutinise a sample of scripts from the previous 3 years' tests at the relevant boundary. They then scrutinise a small number of packs of scripts from the live test, where each pack corresponds to one of a range of marks around the draft threshold. By judging which packs contain	The use of a committee of high status subject specialists - the senior external markers who were involved in the test development process - enables level boundary recommendations to be determined on the basis of the test-level judgement of experts. (The process assumes that subject experts are able to match test-level performances when the test performances reflect responses to different questions from one year to the next.) It requires decisions to be made through group consensus.

	scripts 'below the historical standard', 'equivalent to the historical standard' or 'above the historical standard', the markers recommend the level boundaries. These judgements are generally conducted in the absence of additional statistical information.	
Level confirmation exercise - **Live Data Collection** (June 2000)	Each marker returns the marks of a randomly selected sample of (up to 22) pupils from one of the schools that she is marking. These data are analysed by QCA statisticians (e.g., the extrapolated proportions of pupils at each level given the draft thresholds; gender breakdowns of performance; etc). This process takes into account the effects of borderline remarking on the final level distributions.	The QCA evaluates the performance of pupils on the live test in order to determine the appropriateness of the draft level threshold extrapolations (which were based upon the performance of pupils on the same test when it was at pre-test stage).
Level confirmation exercise - **Level Threshold Setting** (June 2000)	The QCA convenes a meeting of representatives from the QCA, the TDT, the EMA/senior markers and independent observers and stakeholders (from the teaching and academic sectors). Each of the draft thresholds are revisited in the light of recommendations from the Script Scrutiny and the Live Data Collection exercise. Final level boundaries are determined.	The group consensus - of representatives of various stakeholders in the test development and awarding process - is used to integrate various sources of information concerning the final location of level boundaries. The committee is empowered to make final level boundary decisions.
Test Evaluation (post June 2000)	QCA commissions independent evaluations of how the final tests functioned.	Independent reviews feed into the design and standardisation of subsequent tests.

Appendix 2: The Nature of Test Potential

The present report proposes that the weak criterion referencing described by Baird *et al.* (2000) re-defines comparability in terms of matching (hypothetical) qualities of candidates rather than (actual) qualities of candidates' performances. The hypothetical quality upon which matching is supposed to be based is the construct of test potential. Test potential can be defined as the performance that might be expected from a pupil on an idealised test, given her cognitive, conative and affective relation to a specific curriculum.[24]

Test potential is, therefore, conceptually distinct from more limited psychological constructs like ability, because it is an amalgam of all the factors that might enhance test performance. Indeed, we might expect test potential to be causally related to numerous pupil-level factors, for example:

- enjoyment of a curriculum area;
- tenacity or motivation to succeed;
- base level of understanding of a curriculum area (i.e., before embarking on a course);
- general level of intelligence.

No doubt, certain of these pupil-level factors would be causally related to school- or system-level factors, for example:

- quality of teaching in a curriculum area;
- school ethos;
- amount of money provided by government for education;
- base levels of understanding of a curriculum area within society;
- changing lifestyles and attitudes to education within society.

While, for example, cohort referencing would 'partial out', or ignore, the effects of such factors upon overall 'pass-rates' from one year to the next, test potential referencing

---

[24] By 'idealised' the definition implies a test of a given level of difficulty.

explicitly takes these effects into account. That is, if pupils were generally better taught in year 2 than year 1, this would have an impact upon 'pass-rates'. Likewise, if pupils were generally more motivated in year 2 than year 1, this would have a similar impact. As such, test potential referencing supports the kind of inferences from aggregated test results that the government requires, in light of its national performance targets.

Test potential referencing may be compatible with the notion of level descriptions, *but only if these are quite general and are not taken to be prescriptive of necessary competencies*. The problem, of course, is that not all candidates at a certain level of test potential will necessarily be equally successful in all learning outcomes - this is embodied in the practical imperative underlying mark aggregation that poor performance in one part of a test can be compensated for by good performance in another. Note how this recommends that tests should sample widely from a curriculum's potential learning outcomes in order reliably to indicate test potential. Level descriptions can be helpful in giving a general impression of the type of performances that might be expected of the 'average' pupil at each level. Such descriptions would be useful for test result users, although not particularly for those responsible for setting and maintaining test standards.

Test potential is a construct that is internal to a specific curriculum. It does not provide a conceptual framework for determining comparability between curriculum areas (e.g., between maths and science at Key Stage 2).[25] More importantly, test potential is not likely even to provide a conceptual framework for ensuring comparability within the same curriculum area when significant curriculum changes occur. This is because test potential, as defined in relation to the original curriculum, may well not relate in the same way to the new curriculum. If it cannot be assumed that pupils would receive similar rankings in idealised tests of the original and new curricula - that is, if the tests measured essentially different and imperfectly correlated qualities - then attempting to match by test potential would not be appropriate. In fact, there may simply be no rigorous conceptual frameworks through which to determine comparability in such situations (see also Newton, 1997b).

---

[25] It could only do so in the situation in which test potential in maths was the same as test potential in science for all members of a population. In practice, tests in different subjects do not correlate that closely. If they did, there would be little need for subject-specific tests at all.

When curriculum change is small from one test to the next, a program for maintaining standards might 'borrow' from the test potential framework. This might be the case if a relatively small part of a curriculum had been replaced by a new part (that required the same amount of curriculum time to be taught and that stretched candidates in similar ways and that was assessed in a similar manner). In this situation, cut-score decisions might be guided by performance in curriculum areas that have not changed. However, where significant curriculum change does occur, and the new test essentially assesses a different (aggregate of) construct(s), the test potential framework cannot support the maintenance of test standards. As Cresswell (1996) notes, in such situations test standards would need to be *re-set* and this could only be done through the value judgements of professionals who were acknowledged to be subject matter experts.

Exactly, what teachers would be asked to do in this 're-setting' situation is less clear, particularly if some form of notional equivalence with previous forms of the test was required (as is often the case). What could it mean for two tests to be comparably graded when they assessed related, but different qualities? Cresswell (1996) argues that the only way ahead is to argue that tacit standards for judging such matters reside as a dynamic norm established within the teaching profession. Wiliam (1996) similarly argues that such tacit standards can be said to reside as shared, inter-subjective, constructs within a community of practice.

As a final point, it is important to note that there are those who would doubt that test potential is a meaningful construct. Davis (1998), for example, might respond by suggesting that humans simply do not possess such enduring traits. While, there may be good philosophical and psychological reasons to sympathise with such a position, the response to this is essentially pragmatic. When high correlation is observed across different tests that are supposed to assess the same construct, then test potential may be considered to be a useful construct. That is, there would appear to be a consistent quality underlying pupils' test performances and it is this quality that we refer to as test potential.

In reality, though, the purist is probably correct in arguing that test potential is not a truly meaningful construct when scrutinised in depth. We know, for instance, that differing question formats elicit different qualities of performance from different sub-groups of pupils (e.g., boys tend to perform better in a curriculum area, relative to girls, if it is

assessed using multiple choice questions). To the extent that the difficulty of a test differs across pupils, the idea of an idealised test - 'a test of given difficulty' - becomes problematic (meaning that test potential, strictly speaking, would need to be operationally defined in terms of a specific question format as well as a specific curriculum). Moreover, we know that individual pupils perform differently in different areas of the curriculum (meaning that test potential, strictly speaking, would need to be operationally defined in terms of tests that appropriately sampled all curriculum areas). No doubt there are many other ways of challenging the idea that a single score can constitute a valid representation of the knowledge, skills and understanding of a pupil in a certain curriculum area.

NFER HEAD OFFICE
National Foundation
for Educational Research
The Mere
Upton Park
Slough
Berks SL1 2DQ.
Tel: 01753 574123
Fax: 01753 691632
E-mail: enquiries@nfer.ac.uk
Web site:http: //www.nfer.ac.uk

NFER WELSH OFFICE
Chestnut House
Tawe Business Village
Phoenix Way
Enterprise Park
Swansea
SA7 9LA.
Tel: 01792 459800
Fax: 01792 797815
E-mail: scyanfer@abertawe.u-net.com

NFER NORTHERN OFFICE
Genesis 4
York Science Park
University Road
Heslington
York
YO10 5DG.
Tel: 01904 433435
Fax: 01904 433436
E-mail: jbh3@york.ac.uk