LEVEL BEST ?

Levels of attainment in national curriculum assessment

REPORT 2001

DYLAN WILIAM, King's College London, University of London

Published by the Association of Teachers and Lecturers

CONTENTS

The Association of Teachers and Lecturers (ATL) is a TUC-affiliated trade union and professional association working for a dynamic, education service which responds to the needs of students and to those who teach them.

ATL provides high-quality professional support to over 150,000 members, who are drawn from every sector of education. Non-party political but politically astute, ATL also plays an active part in shaping Government policy and responding with authority to consultations. Decision-makers turn first to ATL because of its constructive approach, which is based on the practical experiences of its members – real practitioners, reporting from the frontline of the profession.

To receive the text of this book in large font, please contact ATL HQ.

ABOUT THE AUTHOR

Dylan Wiliam is Assistant Principal and Professor of Educational Assessment at King's College London. After working as a teacher for eight years in inner-city schools in London, he joined Chelsea College as Nuffield Research Fellow on the Graded Assessment in Mathematics project. After the merger of Chelsea College with King's College he became Lecturer in Mathematics Education at King's. In 1989 he was appointed as Academic Co-ordinator of the Consortium for Assessment and Testing in Schools which developed a variety of statutory and non-statutory assessments for the national curriculum of England and Wales – described by Kenneth Clarke, then Secretary of State for Education, as 'elaborate nonsense'.

He is consulted regularly by Government agencies on matters relating to assessment and evaluation and he has published over 100 articles, book chapters and books in mathematics education, education law and educational assessment.

His teaching on Masters and Doctorate programmes includes courses on educational assessment, research methods, and the use of information technology in academic research. He divides his research time between mathematics education and research in educational assessment and evaluation, where his main interests are the interplay between meanings and consequences in conceptualisations of validity, and formative assessment.

Outside work, he spends as much time as he can watching baseball, reading and scuba diving.

INTRODUCTION

It is now over 13 years since the Task Group on Assessment and Testing (TGAT) published its proposals for a framework for the reporting of national curriculum assessment results. Given the developments that have taken place over the intervening years, it is remarkable how well this document stands up to re-reading today. However, while the principles underpinning the report are still sound, the recommendations – perhaps inevitably – were not implemented in a coherent way. As the context of education policy has changed, these incoherencies have become magnified. It has become clear that policy-makers have never really understood the nature of the levels of attainment in the national curriculum. The result is that current policy has driven teachers and schools in ways that are antithetical to high-quality teaching and learning.

In some senses this is understandable, as the rationale for the TGAT proposals and the empirical evidence underpinning them has never been fully spelled out. Given the uses that are currently being made of national curriculum test results, it is, however, essential to understand what national curriculum assessment can – and more importantly, cannot – deliver.

Written for both practitioners and policy-makers, the purpose of this book is to explain:

- why the TGAT proposals were framed in the way they were
- the differences between what was proposed and what was implemented
- the limitations of the current system, and what can be done about them.

BACKGROUND

The Government announced its intention to introduce a national curriculum for all students of compulsory school age in England and Wales in January 1987. The formal proposals were published in July that same year [1]. The curriculum would be defined in terms of the 'foundation' subjects that would have to be taught to all pupils of compulsory school age (5 to 16 years), and for each subject, the curriculum would be specified in terms of attainment targets and programmes of study.

The role of attainment targets was to:

establish what children should normally be expected to know, understand and be able to do at around the ages of 7, 11, 14 and 16 and will enable the progress of each child to be measured against established national standards. They will reflect what pupils must achieve to progress in their education and to become thinking and informed people. The range of attainment targets should cater for the full ability range and be sufficiently challenging at all levels to raise expectations, particularly of pupils of middling achievement, who are frequently not challenged enough, as well as stretching and stimulating the most able (pages 9-10).

Alongside the attainment targets, programmes of study would be specified for each subject. These would:

reflect the attainment targets, and set out the overall content, knowledge, skills and processes relevant to today's needs which pupils should be taught in order to achieve them. They should also specify in more detail a minimum of common content, which all pupils should be taught, and set out any areas of learning in other subjects or themes that should be covered in each stage (page 10).

The achievements of each student in terms of the attainment targets were to be assessed at the ages of 7, 11, 14 and 16 (the end of each of the four 'key stages' of compulsory schooling), and reported to parents. In addition, the aggregated results for each school would be made public to provide a performance indicator of the quality of educational provision within the school.

It was proposed that much of the assessment at the ages of 7, 11 and 14 would:

be done by teachers as an integral part of normal classroom work. But at the heart of the assessment process there will be nationally prescribed tests done by all pupils to supplement the individual teachers' assessments. Teachers will mark and administer these, but their marking – and their assessments overall – will be externally moderated (page 11).

The existing school-leaving examination, the General Certificate of Secondary Education (GCSE) would remain as the predominant means of assessment at the age of 16.

The consultation document also proposed setting up a Task Group on Assessment and Testing (TGAT) – an interesting title, since it demonstrates that at the time, the Government regarded assessment and testing as somehow different activities. As defined in the July 1987 consultation document, the brief of the Task Group was:

to advise them [ie, the Secretary of State for Education and Science and the Secretary of State for Wales] on the overriding requirements that should govern assessment, including testing, across the foundation subjects and for all ages and abilities, with a view to securing arrangements which are simple to use and understand for all concerned, helpful to teachers and appropriate for the purposes of assessment [...] and affordable (page 26).

When it was set up in September 1987, the Task Group was given four terms of reference, of which the first was the most important.

To advise the Secretary of State on the practical considerations which should govern all assessment including testing of attainment at age (approximately) 7, 11, 14 and 16, within a national curriculum; including

- the marking scale or scales and kinds of assessment including testing to be used,
- the need to differentiate so that assessment can promote learning across a range of abilities,
- the relative roles of informative and of diagnostic assessment,
- the uses to which the results of assessment should be put,
- the moderation requirements needed to secure credibility for assessments, and the publication and other services needed to support the system –

with a view to securing assessment and testing arrangements which are simple to administer, understandable by all in and outside the education service, cost-effective, and supportive of learning in schools [2 (appendix A)].

The Task Group published their recommendations in January 1988. The group also produced three supplementary reports and a digest for schools, which were published later in the same year [3,4].

WHAT DID THE TGAT REPORT SAY?

The Task Group identified four purposes that the information derived from national curriculum assessments should be capable of serving:

formative, so that the positive achievements of a pupil may be recognised and discussed and the appropriate next steps may be planned;

diagnostic, through which learning difficulties may be scrutinised and classified so that the appropriate remedial help and guidance can be provided;

summative, for the recording of the overall achievement of a pupil in a systematic way;

evaluative, by means of which some aspects of the work of a school, an LEA or other discrete part of the educational service can be assessed and/or reported on (paragraph 23).

The report argued that formative and diagnostic assessments can be aggregated to serve summative and evaluative uses, but that the reverse process (disaggregating data from assessments derived for summative and evaluative purposes) is impossible. For this reason, the group recommended that:

an assessment system designed for formative purposes can meet all the needs of national assessment at ages before 16. [...] Summative judgements or aggregations may be made at ages other than 16, but only at 16 does it seem appropriate for assessment components to be designed specifically for summative purposes (paragraph 26).

The descriptions of formative and diagnostic purposes of assessment given above overlap considerably and this is conceded in the report:

We do not see the boundary between the formative and diagnostic purposes as being sharp or clear. If an assessment designed in formative terms is well matched to the pupil, it is likely to provide some information which will help in the diagnosis of strengths and weaknesses. Some assessments of this kind will be of value as indicators of learning problems which require further investigation. Further diagnostic investigations would, however, often involve the use of highly specialised tests of relevance to only a limited set of circumstances and pupils (paragraph 27).

In order to support learning, the information from national curriculum assessments would of course have to be fed back to the student, but the Task Group also said that it should also be 'fed forward' to the student's future teachers. To ensure that this information is as useful as possible to teachers, the Task Group recommended the use of 'profile components'.

Profile components are not defined in the report but are characterised by examples. The examples offered for science are derived from the Graded Assessment in Science Project [5], which assesses and reports science attainment in terms of skills, knowledge and understanding, and exploration. The report also gives an example for English which, it is suggested, should be profiled in terms of writing, oracy, reading comprehension, and listening.

It is clear therefore that the term 'profile component' was used to describe a reporting component which must relate to a clear and widely understood construct. The report further recommended that:

an individual subject should report a small number (preferably no more than four and never more than six) profile components reflecting the variety of knowledge, skills and understanding to which the subject gives rise. Wherever possible, one or more components should have more general application across the curriculum: for these a single common specification should be adopted in each of the subjects concerned (paragraph 35).

Having determined that attainment should be reported in terms of profile components, the next step for the Task Group was to determine how the attainment on each profile component was to be reported. The Government's view was apparently that the primary function of the assessment was that of monitoring whether students had made the progress that would be expected from a student of their age. In the original consultation document for example, attainment targets were defined as establishing

'what children should *normally* [my emphasis] be expected to know, understand and be able to do at around the ages of 7, 11, 14 and 16' (page 9).

The implication is clearly that the attainment targets were to provide 'benchmarks' [6 (page 50)]: performance standards against which a student could be measured and found either 'satisfactory' or 'wanting'.

The difficulty with such simple benchmarks is that if they are sufficiently demanding to provide real challenges for the most able, then they are so far beyond the reach of most students that these students are likely to give up. Conversely, if they are set so as to be motivating for the lower attainers, then they provide no motivation for the higher attainers, who will quickly see that a 'satisfactory' score can be attained with little effort.

This difficulty may have been realised by the authors of the consultation document, because the attainment targets were required to be differentiated in some sense:

the range of attainment targets should cater for the full ability range and be sufficiently challenging **at all levels** [emphasis in original] to raise expectations, particularly of pupils of middling achievement who frequently are not challenged enough, as well as stretching and stimulating the most able (page 10).

However, this merely compounds the difficulty. If there are a variety of benchmarks for each age group, how is anyone to know which benchmark is the 'right' one for a student? The consultation document states that HMI reports show that

'a weakness far too frequently apparent in the present system is under-expectation by teachers of what their pupils can achieve' (page 3).

If such under-expectation were prevalent, then it would certainly be perpetuated by allowing a variety of benchmarks, because it is likely that students would be entered for the 'wrong' benchmarks – ie, those that were too easily achieved.

A system of multiple benchmarks – in other words, some sort of scale – might therefore not combat under-expectation, but it seems likely that such a system would be less demotivating than a single benchmark for each age group.

The Task Group considered two main approaches. The first was that the reporting structure should consist of separate sets of benchmarks for each of the four age groups, so that those for 7 year-olds would apply only to 7 year-olds, those for 11 year-olds only to 11 year-olds and so on. Crucially, the four sets of benchmarks would be independent of each other. The second approach was that a common set of benchmarks should apply across the age and achievement range so that, in particular, the reported scores at any one key stage would directly relate to those at other key stages.

WHY NOT USE AGE-SPECIFIC LEVELS OF ATTAINMENT?

Any framework which used independent reporting scales at each of the reporting ages would, in effect, be using age-specific levels of attainment. In general, there is no need for such a structure to have the same number of levels or grades at different reporting ages, or for the grades reported at different ages to be comparable in any sense.

It is also not necessary for the levels to be 'equally-spaced' in the sense that the system is designed so that equal numbers of students achieve each level. Indeed, as the TGAT report points out (paragraph 97) it is commonly the case that the scores are arranged so that comparatively few students are awarded those scores at the extremes.

As is clear from the interim report of the Science working group, the Government originally seemed to want just three levels of attainment (for 'below-average', 'average', and 'above average' students) at each of the ages of 7, 11 and 14, perhaps graded A, B and C, with the seven grades of the soon to be introduced GCSE at age 16 [7 (page 36)]. The TGAT report discusses the advantages of such systems for the reporting of student attainment in terms of their 'apparent uniformity across all ages', but notes that such scales also have drawbacks.

One difficulty cited by the report is the possibility that students will make progress in absolute terms between reporting points, but their reported score would appear to be going down. For example, a pupil might be reported at grade A at age 7, B at age 11 and C at age 14, despite having made steady progress in absolute terms.

Even students whose attainment is progressing satisfactorily would be reported at the same point (eg, grade B) each time they were assessed, as the report notes. The intended interpretation of this is that students are progressing satisfactorily – and such a framework of scales might be appropriate for recording features such as effort – but many students would be disappointed to find that, despite having made significant gains in attainment, their reported grade had remained unchanged. Indeed, the work of American psychologist Carol Dweck has shown that such feedback can have highly negative consequences for learners.

In a series of research projects over many years, Dweck and her colleagues have shown that most students hold one of two views about the nature of ability [8]. This is important because the view of the nature of ability has a profound impact on how students react to challenging tasks. Some students see ability as a fixed entity – broadly, how clever you are is how clever you stay. Viewing ability as fixed, their response to a challenging task depends upon their belief in their chances of success in the task. If they believe their chances of success are high, then they will engage in the task in order to have their ability re-affirmed. If however they believe their chances of success are low, they are likely not to engage in the task, in order to avoid being 'shown up'. Put bluntly, they – like most of us – would rather be thought of as lazy than thought of as stupid.

Other students see ability as incremental. A challenging task therefore offers a chance to 'get cleverer', and whether their belief in their chances of success are low or high, they will engage in the task in order to improve their ability.

In order to optimise the conditions for learning, it is therefore necessary for students to believe that ability is incremental, rather than fixed. A system of age-dependent levels would lead to a situation in which many students would get the same grade or level at ages 7, 11 and 14, thus potentially reinforcing a belief in ability as being fixed. If (as most people had assumed) there would be more grades or levels for reporting the achievement of older students, the reported attainment of lower-attaining students could actually decline as they became older, thus increasing alienation and disaffection.

In order to minimise these effects – particularly among lower attainers – there was a great deal of interest in the development of graded assessment schemes in the early 1980s. The work of the graded assessment projects supported by the Inner London Education Authority (ILEA) was particularly influential in the development of the national curriculum levels structure.

WHY USE AGE-INDEPENDENT LEVELS OF ACHIEVEMENT?

In the late 1970s and early 1980s, a number of schemes of graded tests in modern languages were developed [9]. The influence of these graded tests can be clearly seen in the recommendations of the Cockcroft report, which urged the development of similar schemes in mathematics [10].

Four teams – for mathematics, science, English and design and technology – had been set up in 1983 to develop systems of graded or graduated assessment. Originally this work focused on upper-secondary school students (age 13-16), but later many of the systems expanded to cover the entire secondary school age range.

A key aim of these teams was to provide a reporting structure that allowed students between the ages of 13 and 16 to experience progression. Rather than thinking of themselves as a 'level X person', they would instead think of themselves as someone who had so far reached level X, but, with more work, would be able to progress to level X+1. For this reason all the teams agreed that, if possible, the levels structure should be constructed in such a way that all students in mainstream schools between the ages of 13 and 16 should have a good chance of achieving one level each year.

In order to work out how many levels this would require, the mathematics team used data from the Department of Education and Science's Assessment of Performance Unit [11-16] and the Concepts in Secondary Mathematics and Science (CSMS) project [17] to produce models of how the proportion of students able to demonstrate particular skills or concepts could increase with age. Of course any model would depend on the emphasis given to various topics and the order in which they were taught, but since the national curriculum in mathematics and in science appeared to be quite similar to the existing curriculum in most schools, this data could be expected to be reasonably robust, at least for these two subjects.

At first, it seemed that between nine and 13 levels would be enough. When detailed simulations of the effects of these levels were carried out though, it became clear that even with 13 levels it would take some students of below-average attainment nearly two years to reach the next level. In the end, it was discovered that the minimum number of levels needed to cover the whole attainment range was 15 [18].

During the autumn of 1987, the Task Group took advice from a number of witnesses (see Appendix C of the report), including Margaret Brown, who was then director of the mathematics team. In particular – given that providing students with a reasonable chance of achieving one level per year would require 15 levels for the 13 to 16 age range – she was asked how many additional levels would be required to extend this down to age 7. Although the evidence in support of this was not as strong as for older students, it appeared that an additional five levels – ie, 20 in all – would be necessary.

While distinguishing meaningfully between 20 levels of achievement might have been possible with mathematics and science, it was not obvious this was possible with design and technology. With subjects like history and English, it was almost certainly impossible. However, the Task Group's brief required reporting only at ages 7, 11, 14 and 16. Therefore, instead of 20 levels with students achieving one level a year, the Task Group proposed a system of 10 levels, with one level being achieved every two years. The decision to propose a system of 10 levels was therefore not just the result of plumping for a nice round number: there is a clear rationale for why there were 10 levels, rather than 9 or 11 or 5 or 20. The ten-level framework was the result of a clear priority to provide a system that allowed students to experience progression (in order to promote a view of attainment as incremental rather than as fixed) and that ensured the focus was on progress. rather than upon absolute levels of achievement.

CAN YOU REALLY 'LEVEL' ATTAINMENTS THIS WAY?

It is sometimes claimed that while the TGAT model might work well for some subjects such as mathematics and science (which are organised hierarchically and involve the accumulation of knowledge) it does not work so well for subjects like English or history [19,20]. While there will be different practical difficulties in implementing levels in different subjects, the principle of the TGAT model does in fact, apply equally well in all subjects (as was re-iterated by the Task Group in the second of its three supplementary reports (paragraphs 7-8)).

With a progressive model, the essential task is to define the nature of progression – in other words, when someone gets better at a subject, what is it that gets better? Once we know this, we can say what it is that a student should do next, given they have reached a particular point in their learning. The answer to this question may be difficult, but to say that there is no answer is to deny the possibility of progress in the subject at all.

In defining progression in a subject, it is important to realise that this is a value-based decision. With science for example, many argue that an essential element of progression is the accumulation of knowledge and facts. With English on the other hand, the definition of progression is much more contested. For some people, getting better at for example writing involves conveying meaning with a growing sense of audience and purpose. For others, progression might be limited to being able to construct more complex sentences and spell a greater number of words correctly. However, once a definition of the nature of progression in a subject is agreed, then it is a straightforward task to define levels of achievement in that subject.

With mathematics, the nature of progression is fairly well understood, at least in comparison with other subjects. The current attainment targets appear to capture this notion of progression fairly well. With design and technology, progress consists of an increasing range of competences and skills, demonstrated across an increasing variety of contexts, with a growing range of materials, and taking into account an increasing number of variables. However, in the current orders, the competences and skills are defined in the attainment targets, while the range of materials is defined in the programmes of study. This leads to a situation where a student can achieve level 3 at key stage 1 by demonstrating a particular level of capability, but fails to do so at key stage 3 because the range of materials with which the competences can be demonstrated has not increased sufficiently. This issue is explored further below.

With history, while we may yet be unclear about what it is that develops when somebody gets better at history, we do at least know that the answer is certainly not chronology. There is no sense in which being 'very good' on the Romans is equivalent to being 'a bit below average' on the Victorians, so attainment targets based on chronology would be doomed to fail.

Certainly few people would disagree that one of the things that develops as a student improves at history is that they have more historical facts and relationships at their disposal. For this reason, knowledge can and should feature within the attainment targets for history. However, if this is so, it is essential to regard the levels as cumulative. If, for example, level 8 (including assessment at level 8) does not subsume the knowledge and experience associated with the preceding levels, then a student would get level 8 for knowledge simply by knowing the facts associated with level 8. Since one fact is no harder to learn by rote than any other, then we should not be surprised to discover that the history curriculum in all schools were dominated by the teaching of 'level 8 facts'.

Robert Skidelsky was right when, in his speech at the Centre for Policy Studies conference [19], he identified the atomisation of attainment targets into over-precise statements of attainment and 'crude criterion referencing' as the cause of the difficulty. He was wrong however when he said that this is a consequence of the TGAT model. The idea of statements of attainment does not in fact feature anywhere in the TGAT report – they were actually the creation of the mathematics and science working groups, which also began their work in September 1987, but who worked independently of the Task Group.

The TGAT model can be applied in any subject where consensus can be found about what students should do next once they have reached a particular stage. Paul Black [21] points out that the statements of attainment can be regarded as *defining* competence at a particular level – an approach which Popham [22] claims has failed in the United States – or *exemplifying* the kind of standard required (the approach recommended by Black).

There are views of education that would lead one to reject the TGAT 8-level model. For example, if one believed that students who had failed to achieve a certain level by a certain age should then work towards a different set of attainment targets than a student who had achieved the same level at a younger age, then one would certainly have to reject the 'common ladder' that this model represents. Such a view is held for example by those who argue for separate and different curricula for students between the ages of 14-19 according to whether the students have an 'academic' or 'vocational' bent.

This approach however does not appear to be widely supported, and most people's views of what the nature of the school curriculum should be are entirely consistent with the 8-level model. Skidelsky's own characterisation of progress seems to summarise this approach rather neatly:

> higher levels of attainment in nearly every subject involve many activities which both select and synthesise numerous aspects of the specific knowledge, skills and understanding which characterise those subjects.

Nothing in the TGAT proposals is inconsistent with this.

DO THE LEVELS RELATE TO STAGES IN INTELLECTUAL DEVELOPMENT?

Once we accept that there is a meaningful notion of progression (that is, we agree on what it is that gets better when a student gets better), then the levels are arbitrary – we can site them where we like. In particular, the levels do not purport to correspond to significant stages in a student's development. The TGAT report suggested that the levels should be equally spaced with respect to time, whereas other schemes – such as the ILEA graded assessment systems – proposed that earlier levels should be closer together than later levels.

What the Task Group actually proposed was that to begin with, the performance requirements of level 2, (for example), would be pitched so that the majority of students would achieve level 2 by the end of key stage 1, but some would achieve level 3, while some would achieve only level 1. At the age of 7 therefore, the median student is not on the borderline between level 1 and level 2, but well on the way (in fact exactly half-way!) to level 3.

It is the median *6 year-old* who is just on the threshold for level 2 (and this was made explicit in the Government's specifications to the consortia developing the first 'standard assessment tasks' for 7 year-olds). Similarly, while the median 11 year-old will achieve level 4, she or he does so comfortably. It is the median 10 year-old who just 'scrapes' a level 4 in the TGAT framework.

As noted above, these definitions are arbitrary. However, once we have defined where each of the levels is situated, we can have no say over what proportion of students at other ages achieve the same level. For example, while level 6 was defined to be the median attainment of 14 year-olds, then how many 11 year-olds will achieve this level is an empirical question, which can be answered only by assessing 11 year-olds and seeing how many of them achieve the same level as median 14 year-olds. The TGAT report did provide some 'rough speculation' about how the range of achievement might vary, and this has, in fact, proved quite accurate [23].

DOESN'T THIS IMPOSE A 'LOCK-STEP' RATE OF PROGRESSION ON STUDENTS?

When the first TGAT report was published, many people incorrectly interpreted the proposals as imposing a 'lock-step' rate of progress on students – in other words that the learning of individual students is somehow neat and linear and that all students will learn at a particular rate. This is, of course, not so – the progress of any student is very complicated, with learning proceeding at different rates in different areas. Nevertheless it does make sense to talk about what the average student at a given age will be able to do, because this need not be the same student at each age. An individual student might be average at the age of 7, but below average by the age of 11. The same student might, between the ages of 11 and 14, experience a 'spurt' of growth in attainment, and at age 14 be well above the average.

Furthermore, not all students will progress at the same rate in different aspects of a subject. For all students there will be a degree of 'trade-off' with strengths in some aspects of a subject and weaknesses in another. Students who have median attainment overall may not be average at anything. They may be above average on some things, and below average on others.

DOESN'T THE WAY NATIONAL CURRICULUM LEVELS ARE DEFINED HAVE MORE TO DO WITH NORM-REFERENCING THAN CRITERION-REFERENCING?

The important distinction between norm-referencing and criterion-referencing (indeed the reason why criterion-referencing was introduced in the late 1960s) is that for norm-referencing, all we need is to be able to put students in rank order – there is no need to understand what we are putting them in rank order *of*. In contrast, assessing in terms of criteria rather than norms requires us to be absolutely explicit about what we are assessing, and thus criterion-referenced assessments have to be explicitly linked to learning outcomes.

Given this, defining levels in terms of the achievements of median students appears to be norm-referenced rather than criterion-referenced, but in fact all criterion-referenced systems must start this way. The meaning of a criterion such as 'can spell simple monosyllabic words correctly' depends on how we interpret the word 'simple' (and also on what we mean by 'monosyllabic'!). The important point however about a criterion-referenced system is that once we have agreed which words are simple and which are not, then that definition does not change, no matter how many students achieve it. The usual example of this is the driving test. The standards of the driving test have been set so as to represent a reasonable standard of safety but at the same time must be achievable by humans. If driving licences were given only to those who could read a standard number-plate at 75 metres rather than the currently required 25 metres, then no one would pass! It was this observation that led William Angoff to remark that if one scratches the surface of any criterion-referenced assessment, one always finds a norm-referenced set of assumptions lurking underneath [24].

CAN YOU USE NATIONAL CURRICULUM ASSESSMENT TO MEASURE STANDARDS OVER TIME?

The definition of levels proposed by the TGAT report means that it is possible (indeed, straightforward) to measure standards over time [25] with one important condition: you must not change the curriculum. In practice, this condition is never satisfied, because curricula always change. For example, there can be little doubt that the ability of British secondary school students to compute with logarithms has declined woefully over the last thirty years – because of course, this is no longer taught.

Another reason why any attempt to measure standards over time is doomed is that what is actually taught in schools changes, even if the 'official' curriculum does not change. For example, the pressure on primary schools to improve their key stage 2 results in English, mathematics and science has increased markedly in recent years. Understandably, schools have therefore focused more closely on these subjects in order that their students' results in these key stage 2 tests improve [26].

If students spend more time learning English, mathematics and science and less time on other subjects, then levels of achievement in English, mathematics and science are likely to rise. By placing greater and greater pressure on schools to improve test results in certain subjects, improvements in these subjects are likely to be secured. This is only at the expense of standards of achievement elsewhere though – particularly in non-tested subjects, but also in other areas such as personal and social development.

The same phenomenon can be observed within subjects. The end-of-key-stage tests can assess only a small proportion of the national curriculum in each tested subject. In low-stakes contexts, the limited range of achievement that is assessed in the tests can stand as a proxy for achievement across the whole subject. However, in high-stakes contexts, there is pressure to increase the student's performance in those aspects of the subject that will be tested. Since there is evidence that teachers can predict which aspects of a subject will be tested [18 (Appendix C)], we should not be surprised that these aspects are given more teaching time. Standards of achievement in the tested areas will rise, but only at the expense of untested areas. Therefore while the *reported* standards of achievement may rise, the actual level of achievement across the whole subject could well be falling, and the tests are no longer an adequate proxy for achievement across the whole domain. Put bluntly, the clearer you are about what you want, the more likely you are to get it, but the less likely it is to mean anything. For all of these reasons, the whole idea of measuring standards over time in any real sense is nonsense – not just with national curriculum tests, but with any tests of achievement.

Interestingly, there *is* evidence that young people are getting cleverer. In every country in which this has been studied, there is evidence that the IQ of the population has been rising steadily over recent decades. While it is not clear whether this is due to better schooling, nutrition, or a greater degree of stimulation in children's early years, average IQs appear to be rising at the rate of approximately one point every three years [27].

Perhaps more importantly, the improvements that we have witnessed in national curriculum test results over the past few years (particularly at key stage 2) may not just be partial, but may actually be counter-productive in terms of meeting the nation's needs. While attempts by the Office for Economic Co-operation and Development (OECD) to establish any kind of link between standards of achievement in schools and industrial productivity have been unsuccessful [28], there is a growing acceptance that employment in the twenty-first century will require a flexible workforce which is able to learn new skills and adapt quickly to new demands. These kinds of skills are precisely those aspects of the curriculum that are currently receiving less emphasis in the drive to improve results in tests which measure only those things that are easy to measure in timed, written tests – what Stephen Ball has called 'the curriculum of the dead' [29].

ARE THE LEVELS COMPARABLE BETWEEN SUBJECTS?

The issue of standards of difficulty between subjects is one of the most misunderstood in assessment. The question of whether standards of achievement are comparable between subjects is quite meaningless, as highlighted by Robert Wood's article in 1976 *Your chemistry equals my French* [30]. Children are learning to speak from the moment they are born (some would say earlier) and so when they take a GCSE in English, they have, in effect, been studying the subject for 16 years. This is somehow meant to be equivalent to taking a GCSE in economics, which they are likely to have been studying for only two years. This illustrates a fundamental feature of standards, which is that they – like criterion-referencing – rest on norm-referenced assumptions.

Imagine, for a moment, that we introduce two new GCSEs – chess and computer games playing. How do we decide what level of performance to require at each grade? The obvious answer (indeed the only answer that makes any sense) is to fix the standards by reference to other subjects. So someone who is average at each of these subjects would achieve a grade C, and we would fix the standards expected at other grades to be broadly consistent with other subjects. Almost all students, with application, would be able to achieve at least a grade G, while the standard for A* would be set so that no more than say, 2% of students would be able to achieve it.

Once we have set standards in this way, consider a student who achieves a grade A in chess, a grade E in computer games playing and C grades in every other subject. We might say that chess was their best subject, but it does not make sense to say that their chess playing is better than their mathematics and their English. The only way than standards can be compared between subjects is by reference to a population. Furthermore, the relationship between subjects changes over time.

Since many more students will have been engaged in computer games playing than in chess during the period prior to the introduction of the two new GCSEs, it is likely that explicit teaching in these new subjects would have a greater impact on performance in chess than in computer games playing. Pretty soon, we might find that the average grade in chess would be a grade B, while the average grade in computer games playing remains a grade C. There might even be public alarm that standards of achievement in computer games playing were lagging behind those in chess because computer games playing skills were felt to be more important to our industrial competitiveness than chess.

The various standards that we associate with different subjects are the result of numerous forces that have shaped their development, and the impact of these forces has been different in different subjects. This is why the question of whether standards of achievement in English are comparable to those in mathematics is not just difficult to answer in practice – it is a question that is fundamentally meaningless, except in a trivial sense that two norm-referenced tests are equally hard because the average scores on the test are the same.

ARE THE LEVELS COMPARABLE BETWEEN DIFFERENT ASPECTS OF THE SAME SUBJECT?

The question of whether levels are comparable between different aspects of the same subject is, on a smaller scale, exactly the same issue as that of comparability between subjects.

Politicians frequently remark that standards of writing are not as high as standards of reading at key stage 2. It *is* possible to say that since the standards in the two aspects of literacy were set, those in reading have improved more than those in writing. It is also possible to say that standards of reading are adequate for a specific purpose (eg, coping with the demands of key stage 3) while those of writing are not (were this the case). This too is circular however, since what we expect students to be able to do at key stage 3 is again the result of the history of our education system. With a different set of beliefs about what ought be happening at key stage 3, it might be that the current standards of achievement are entirely adequate.

ARE THE LEVELS COMPARABLE BETWEEN KEY STAGES?

This is probably the key question in the current debate, because the results of students at the end of the key stages are being used as if the levels from different stages are comparable. However, as is made clear below, this usage of results is based on a failure to appreciate the significance of what happened after the TGAT report was published.

The original proposals of the TGAT report clearly intended that the levels should be comparable between key stages – indeed, this was the entire rationale of the reporting system proposed. However, when the Task Group completed its work in December 1987, the only subject working groups that had been set up were those for mathematics and science, and these had produced only draft reports. The legislation that would be necessary to implement the proposals was still under debate in Parliament, and would not receive royal assent for a further seven months. This is significant because at the time that the Task Group's proposals were made, the exact form that attainment targets and programmes of study would take had yet to be determined.

When they were finally published 18 months later, both curricula presented the attainment targets as a large number of 'statements of attainment' (297 in mathematics, 407 in science). This atomisation of the curriculum was not in the TGAT proposals, but emerged from the work of the mathematics and science working groups, and followed quite closely the approaches that had been taken by the graded assessment teams in these subjects.

More significantly for the development of levels, the national curricula for mathematics and science each took radically different approaches to the definition of the programmes of study.

Why programmes of study were included in the definition of the national curriculum at all has never been made clear. After all, if one specifies in detail the material on which students are to be assessed (attainment targets) and how they are to be assessed (assessment arrangements), then what is the point of also specifying what they are to study? The most plausible reason appears to be that those concerned with drafting the 1988 Education Reform Act assumed that the programmes of study were to take the role of the syllabus in traditional examining, without realising that the attainment targets also fulfilled this role.

In the original version of the national curriculum for science, the programmes of study were presented in the now familiar form of one programme of study for each key stage, while in the first version of the national curriculum for mathematics, there was a different programme of study for each level. There is little doubt that the science model was perceived as being more straightforward to implement – put simply, the programme of study which schools were required to teach students depended only on their ages, and changed only at the end of the key stage. In mathematics, the programme of study to which students were entitled depended on the attainment of each student. Consequently different students in the same class might be entitled to different programmes of study. While difficult to implement, this emphasised the need for curricular differentiation, with teachers required to recognise the different needs of each student.

This issue came to the fore in April 1993, when the then Secretary of State for Education, John Patten, asked Sir Ron Dearing (then Chair of the School Curriculum and Assessment Authority (SCAA)) to undertake a review of the national curriculum and assessment framework. The remit for the review suggested four key areas for review, one of which was 'What is the future of the ten-level scale for graduating children's attainments?' An interim report was published in July 1993, and the final report appeared six months later, in December [31,32].

Both reports betray a profound lack of understanding of the ten-level scale proposed by TGAT. For example, the interim report reiterates the (incorrect) belief that the scale works on an assumption that 'a pupil's development in every national curriculum subject progresses in an orderly way through the ten levels' (page 40) and that it is more suitable for some subjects than others (although there appears to be a recantation of this view in the final report).

The interim report identified eight problems with the ten-level scale – five that SCAA thought might be resolved or mitigated, and three which it suggested were inherent. However, seven of the eight issues were not matters pertaining to the adequacy or appropriateness of the scale itself. They were actually concerned with how the scale had been put into operation. The only substantive issue of viability was the unresolved nature of the relationship between programmes of study and attainment targets.

After much debate, the final report seems actually to prefer the idea of separate end-of-key stage scales (a sort of mini-GCSE at the end of each key stage), but viewed the introduction of such a system by 1995 as too risky.

The Dearing review therefore concluded that the existing common scale should be retained, but that it should be terminated when a student reached the age of 14, with the existing grades of GCSE being used at age 16. There was also a clear recommendation that the attainment targets should form the cornerstone of planning and assessment:

the structuring of progression within each attainment target in terms of ten discrete levels ought to facilitate curriculum planning, the matching of work to pupils of different abilities, and the assessment of pupil progress (paragraph 7.24).

In addition, the final report makes clear that there should also be a programme of study for each key stage, rather than for each level. In this sense, the Dearing review ducked the issue of the relationship between programmes of study and attainment targets. The idea of the ten-level scale survived, but the comparability of levels between key stages (so necessary to ensure progression) was lost. What is more disturbing – given the importance accorded to attainment targets in the Dearing review – was that within a year, SCAA itself announced that:

'We have concluded that it is the programmes of study which should guide the planning, teaching and day-to-day assessment of pupils' work. The essential function of level descriptions is to assist in the making of summary judgements about pupils' achievements as a basis of reporting at the end of a key stage' [33].

The final nail in the coffin was the decision by SCAA to instruct its test developers to base tests on the programmes of study, rather than on the attainment targets.

We therefore have key stage 2 tests that are based on the key stage 2 programmes of study and key stage 3 tests based on the key stage 3 programmes of study. The fact that these programmes of study are different means that comparing the results at key stage 3 with those at key stage 2 makes no more sense than saying that a student must have got worse because they achieved a grade C at GCSE and then a grade E at A level in a particular subject.

CAN LEVELS BE USED AS MEASURES OF SUCCESS AND FAILURE?

Data published by the Qualifications and Curriculum Authority (QCA) suggests that around 10% of students achieve the same level at key stage 3 as they did at key stage 2. This has been widely interpreted (including by David Blunkett during his tenure as the Secretary of State for Education and Employment) as an indication that these 10% of students have failed to make progress in key stage 3. This is quite simply not so, and betrays an ignorance of the meanings of the levels as defined in national curriculum, and the limitations of educational assessments.

Firstly, it could be that the students were only just above the threshold of a level at the end of key stage 2 and only just below the threshold of the next level at the end of key stage 3. In other words, they could have made almost a whole level progress (which would be two years' progress for an average child), but the coarseness of the scale does not show the progress made. While two years' progress in three is disappointing, it *is* progress.

Furthermore, the TGAT framework assumes that low-attainers progress more slowly than higher attainers (this was why the graded assessment schemes had the earlier levels closer together). While those in the upper 10% will achieve a level every four terms, the lowest-attaining 10% will take more than three years to achieve a level [23]. In other words, for approximately half of the lowest attaining 10% of students, not changing a level between key stage 2 and key stage 3 is exactly what we would expect if they were progressing steadily.

Secondly, we need to take into account the fact that no assessment is perfectly reliable. Students have good and bad days, markers might be lenient or severe, and the inclusion of particular items in the test will suit some students better than others. The Government has published only limited reliability data on national curriculum tests, but it is likely that the proportion of students awarded a level higher or lower than they should be because of the unreliability of the tests is at least 30% at key stage 2 and may be as high as 40% at key stage 3 [34]. Although this is unfortunate for students who are awarded levels lower than they deserve, these mis-classifications will even out over a class of students – for every student who is awarded a level lower than they should be, there will be another who is awarded a level higher than they should be. However, when we look at specific groups of students – such as the lower-attaining students – then an effect called 'regression to the mean' becomes important. In particular, when looking at the 15% of students (ie, half of 30%) who were awarded a level higher than they should have been at key stage 2, we find that they are unlikely to be as lucky at key stage 3. Consequently, probability dictates that most of them will be awarded the level they should have been awarded at key stage 3. However, some (about 20%) will be unlucky enough to be awarded a level lower than they should have been at key stage 3, and will thus appear to be 'standing still' or even getting worse.

At the other extreme, the opposite will happen. There will be students who appear to have made huge progress in key stage 3, because they were unlucky at key stage 2 and were awarded a level lower than they should have been, but then were awarded a level higher than they should have been at key stage 3. Such students will appear to have progressed three or even four levels during key stage 3, whereas in fact – although obscured by the unreliability of the tests – their real progress could be entirely as expected. It is a well-known fact of educational measurement that change scores (ie, the improvement in scores from one testing occasion to another) are far less reliable than the individual test scores that are being compared.

Thirdly, as noted above, the tests are based on programmes of study which change from key stage to key stage. A student who is awarded for example level 5 on the science test at key stage 2 would not be awarded level 5 if they were to take the key stage 3 science test, because the key stage 3 tests are based on a much broader programme of study. The 'optional' tests that will be used in year 7 and in year 8 will cloud the picture even further, because these will be based on the sequencing of material assumed in the key stage 3 strategy. There is currently no legal requirement to teach particular material from a key stage in a particular year – all that is required is that the programme of study for the key stage must be taught by the end of that key stage. But the year 7 'optional' tests will be based on the material for year 7 of the key stage 3 strategy, and those for year 8, on year 8 of the key stage 3 strategy. So worse than expected scores on the 'optional' tests might just mean that a school has, perfectly legally, not covered the curriculum in the order prescribed in the key stage 3 strategy. In order to avoid this problem, it seems likely that schools will be under pressure to produce 'acceptable' scores in the 'optional' tests so that, unlike TGAT, the key stage 3 strategy *will* impose a 'lock-step' model of progression on the curriculum.

Finally, there will be the effects of teaching to the test, which will be far greater at key stage 2 than at key stage 3. Key stage 2 results are used to compile 'league tables' for primary and junior schools, and so there is a pressure to make these results as good as possible. The same is not true for key stage 3. Indeed, despite the Government's recently announced targets for key stage 3, there is little incentive for schools to teach to the key stage 3 tests, because if they do, they will appear to show less 'value added' at key stage 4. Since many secondary schools regard the key stage 3 tests as a distraction from improving GCSE grades (which are used as *their* 'league tables'), preparation for key stage 3 tests is given far less emphasis than is the case at key stage 2.

Furthermore, as demonstrated above, the key stage 2 tests are based on a much narrower curriculum than those for key stage 3, so teaching to the test is possible. In recent years this has been compounded by the provision of additional money for 'booster classes'.

Taken together, these deficiencies in national curriculum tests mean that the results of key stage tests are almost useless as a measure of success and failure, even if value-added measures were used in place of raw scores.

SUMMARY

The TGAT report proposed a framework for reporting achievement that supported learning. This is because it was based on the notion of ability as incremental rather than fixed, so that the assessment system can be seen as a way of supporting learning rather than grading students on progressively finer and finer scales. The TGAT framework helps to promote the idea that progress depends on factors within the learner's control (such as effort) rather than those outside the learner's control (such as traditional views of IQ). The message to learners in such a framework is 'if at first you don't succeed, try again'.

However, the TGAT proposals were never implemented. Of course the proposals were not rejected, but merely implemented imperfectly.

The first error was not to understand that the key concept behind the TGAT report was that a student's entitlement should depend on their achievement rather than on their age. The second was to compound this by deciding to base the tests not on the attainment targets, but on the programmes of study. The third error was to misunderstand the nature of standards, and in particular the fact that they are fundamentally arbitrary, both between subjects and within subjects. This has led the Government to drive policy in directions that are simply not supported by evidence.

The fourth error was to fail to appreciate the impact of test unreliability on the reliability of change scores for individuals, leading to a mistaken view about what the problem really was. Although it may well be the case that many students fail to progress adequately in key stage 3, (and there is some evidence of this) one cannot deduce this from an analysis of national curriculum test scores. The fifth and perhaps most serious error was to create absolutely irresistible pressures for teachers to teach towards the tests, particularly at key stage 2.

Successive governments have created a situation where policy is based on assumptions that are not merely unsupported by evidence, but just plain wrong. Rising test scores demonstrate little more than teachers' increasing abilities to teach to the tests, and to the power of high-stakes tests to distort the curriculum.

None of this should come as a surprise to the Government. The agencies involved with the development and implementation of the national curriculum and its assessment were warned about these problems 10 years ago. It is not clear whether they kept silent about them, or whether they communicated these concerns to ministers who chose to ignore them. What is clear is that a Government which was serious about raising standards of achievement in schools (as opposed to test scores) would not continue to pursue such incoherent and misguided policies.

WHAT CAN BE DONE?

If the Government is serious about improving standards of achievement in schools, the first step is to ensure that the levels are comparable between key stages. The reason this is so important is that to say a student has demonstrated level 6 on key stage 2 science content really satisfies only the summative function of assessment – in other words, it tells us that this is an able student who has done really well on the key stage 2 science content. It is really only about grading schools and students. Such a result does not tell the teacher of the same student at key stage 3 what that teacher needs to know, which is: 'What, of the things I plan to teach students this year, does this student already know?' That is why the end-of-key stage assessments should be based on the attainment targets rather than on the programmes of study, and the requirements of a particular level should not depend on the age of the student.

In subjects that use tiered tests, this could be done very easily. Instead of having two different tests to assess, for example levels 3-5 (one at key stage 2 and one at key stage 3), there would just be one (this approach has already been explored in Wales). In other subjects not using tiered tests, this would be slightly more complex to implement, but the principle is quite straightforward. The effect of this would probably be to reduce the range of levels achieved at each key stage – particularly in English and science – but the information being passed to teachers at the next key stage would provide clear evidence of where the students were in their learning.

The second step is to ensure that the limited reliability of educational assessments and the impact on national curriculum assessments is understood by all – especially teachers, parents, and policy-makers. A necessary pre-requisite for this will, of course, be undertaking and publishing comprehensive analyses of the reliability of the tests currently in use. This will undoubtedly be politically difficult, since it is likely that the proportion of inaccurate levels awarded in national curriculum assessment is substantially higher than parents or policy-makers believe [34]. Nevertheless, it must be done.

The third (and most difficult) step is to minimise the narrowing of the curriculum caused by 'teaching to the test'. In order to achieve respectable (although not, it has been argued, adequate) levels of reliability, the content of the tests has been restricted to those aspects of each subject that are easy to test reliably (although even with these narrowly focused tests, to get the proportion of students mis-classified at key stage 2 down to under 10% would probably require around 30 hours of testing for each subject, and even longer at key stage 3 [34]). Because the tests assess only one part of the national curriculum in each subject, teaching to the test allows teachers to improve their students' test scores without improving average levels of real achievement. To avoid this, the basis of the assessment must be broadened. The simplest way to achieve this is to is to make greater use of teacher assessment.

Teachers' own assessments of their students are highly reliable, because they are based on hundreds of hours of assessment. These assessments may however suffer from systematic bias. To guard against this, national curriculum tests can be used to moderate the assessments from different teachers. Of course, if the tests are narrow, then teaching to the test would still be a problem, so instead, there could be a large range of tests and tasks, with each student randomly assigned to take one of these tests or tasks. The score gained by an individual student would not provide an accurate measure of that student's achievement, because of the limited reliability of the task or test. But for every student at a school who was lucky in the task or test they were assigned, there would be one who was unlucky, and so the average achievement across the class in the tests or tasks would be a very reliable measure of the real average achievement. Furthermore, the breadth of the tests and tasks would mean that it would be impossible to teach towards the test. More precisely, the only effective way to teach towards the test would be to raise the standard of all the students on all the tests and tasks, which – provided they are a broad representation of the desired curriculum – would be exactly what is required. The Government would have undistorted information about the real levels of achievement in our schools, users of assessment results would have accurate indications of the real levels of achievement of individual students, and teachers would have information about students that would inform their teaching. Isn't that what we all want?

GLOSSARY

criterion-referencing criterion-referencing involves making sense of the result of an assessment in terms of how well the student has achieved clearly defined learning objectives, as opposed to norm-referencing (see below), in which what exactly is being assessed is not explicit.

median the middle value of a set of values when they are arranged in order (or if there is an even number of values, the average of the middle pair). The median is often used instead of the mean (average) because the mean is affected by extreme values while the median is not. For example, if the bosses of a company awards themselves but nobody else a large pay rise, the mean (ie, average) salary of employees at the company will rise, but the median will stay the same.

norm-referencing norm-referencing involves making sense of the result of an assessment in comparison with the attainments of others. For true norm-referencing, this group must have been assessed at some point in the past. Where the score is compared with those by individuals who took the test at the same time, this is better described as cohort-referencing. In simple terms, if sabotaging your neighbour's performance improves your result, then this is cohort-referencing, rather than norm-referencing. The problem with norm-referencing is that all that is required is that you can put students in rank order without have any clear idea of what you are putting them in rank order *of*.

standard the term 'standard' is used to refer either to the level of performance that must be reached, or the number or proportion of some population reaching some specified level of performance. It is this duality of meaning that allows politicians to say that standards are going down when the proportion of students achieving five good grades at GCSE goes down (because the overall level of achievement is lower) *and* when the proportion goes down (because, they say, the exams are getting easier).

REFERENCES

1 Department of Education and Science and Welsh Office, *The National Curriculum 5-16: a consultation document*. 1987, London: Department of Education and Science.

2 National Curriculum Task Group on Assessment and Testing, *A report*. 1988, London: Department of Education and Science.

3 National curriculum task group on assessment and testing, *Three supplementary reports*. 1988, London: Department of Education and Science.

4 National curriculum task group on assessment and testing, *A digest for schools*. 1988, London: Department of Education and Science.

5 Graded Assessment in Science Project, *Trial materials*. ed. JLR Swain. 1987, London: Chelsea College Centre for Science and Mathematics Education.

6 Nuttall, DL, *National assessment: complacency or misinterpretation?*, in *The Educational Reform Act: choice and control*, D Lawton, Editor. 1989, Hodder & Stoughton: London. p. 44-66.

7 National curriculum science working group, *Interim report*. 1987, London: Department of Education and Science.

8 Dweck, CS, *Motivational processes affecting learning*. American Psychologist (Special issue: Psychological science and education), 1986. 41(10): p. 1040-1048.

9 Harrison, A, *Review of graded tests*. Schools Council examinations bulletin, Vol. 41. 1982, London: Methuen.

10 Committee of Inquiry into the Teaching of Mathematics in Schools, *Report: mathematics counts*. 1982, London: Her Majesty's Stationery Office.

11 Foxman, DD, et al., *Mathematical development: primary survey report no 1*. Mathematical development. 1980, London: Her Majesty's Stationery Office.

12 Foxman, DD, et al., *Mathematical development: secondary survey report no 1*. Mathematical development. 1980, London: Her Majesty's Stationery Office.

13 Foxman, DD, MJ Cresswell, and ME Badger, *Mathematical development: primary survey report no 2*. Mathematical development. 1981, London: Her Majesty's Stationery Office.

14 Foxman, DD, et al., *Mathematical development: secondary survey report no 2*. Mathematical development. 1981, London: Her Majesty's Stationery Office.

15 Foxman, DD, et al., *Mathematical development: primary survey report no 3*. Mathematical development. 1982, London: Her Majesty's Stationery Office.

16 Foxman, DD, RM Martini, and P Mitchell, *Mathematical development: secondary survey report no 3*. Mathematical development. 1982, London: Her Majesty's Stationery Office.

17 Hart, KM, ed. *Children's understanding of mathematics: 11-16*. 1981, John Murray: London.

18 Wiliam, D, *Technical issues in the development and implementation of a system of criterion-referenced age-independent levels of attainment in the National Curriculum of England and Wales*. 1993, King's College University of London:

19 Skidelsky, R, *Government testing and the national curriculum – reforming the reformers*. Centre for Policy Studies conference, 21 Sept 1993, Social Market Foundation: London.

20 Marenbon, J, *Testing time: the Dearing review and the future of the National Curriculum*. 1993, London: Centre for Policy Studies.

21 Black, PJ, *The shifting scenery of the National Curriculum,* in *Assessing the National Curriculum*, P O'Hear and J White, Editor. 1993, Paul Chapman Publishing: London. p. 57-69.

22 Popham, WJ, *The instructional consequences of criterion-referenced clarity*. Educational Measurement: Issues and Practice, 1994. 13(4): p. 15-18, 30.

23 Wiliam, D, *Special needs and the distribution of attainment in the national curriculum*. British Journal of Educational Psychology, 1992. 62: p. 397-403.

24 Angoff, WH, *Criterion-referencing, norm-referencing and the SAT*. College Board Review, 1974. 92 (Summer): p. 2-5, 21.

25 Independent Scrutiny Panel on the 1999 Key Stage 2 National Curriculum Tests in English and Mathematics, *Report: weighing the baby*. 1999, London: Department for Education and Employment.

26 Reay, D and D Wiliam, *I'll be a nothing: structure, agency and the construction of identity through assessment*. British Educational Research Journal, 1999. 25(3): p. 343-354.

27 Neisser, U, ed. *The rising curve: long-term gains in IQ and related measures*. APA Science Volumes, 1998, American Psychological Association: Washington, DC.

28 Centre for Educational Research and Innovation, *Human capital investment: an international comparison*. 1998, Paris, France: Organisation for Economic Co-operation and Development.

29 Ball, SJ, *Education, Majorism and the 'curriculum of the dead'*. Curriculum Studies, 1993. 1(2): p. 195-214.

30 Wood, R, *Your chemistry equals my French*, in Times Educational Supplement. 30 July 1976, London. (Reprinted in Wood, R, *Measurement and assessment in education and psychology*. 1987, London: Falmer).

31 Dearing, R, *The National Curriculum and its assessment: interim repor*t. 1993, London: National Curriculum Council & School Examinations and Assessment Council.

32 Dearing, R, *The National Curriculum and its assessment: final report*. 1994, London: School Curriculum and Assessment Authority.

33 School Curriculum and Assessment Authority, *Science in the national curriculum: draft proposals*. 1994, London: School Curriculum and Assessment Authority.

34 Wiliam, D, *Reliability, validity and all that jazz*. Education 3-13, 2001. 29(3): p. 9-13.

ATL 2001
ATL members FREE
Non members £9.99

ISBN 1-902466-07-1

Level Best?

Levels of attainment in national curriculum assessment

It is over 13 years since the report of the Task Group on Assessment and Testing (TGAT) published proposals for reporting the results of national curriculum assessments. Though the principles underpinning the report remain sound, the recommendations were not implemented in a coherent way. Over the years, these incoherencies have become magnified.

Increasingly, national curriculum test results are used to make important judgements about pupils – and their teachers. It is therefore essential to understand what the test results, as measured by the 'levels', can – and cannot – deliver.

Level Best? Levels of attainment in national curriculum assessment explains to both practitioners and policy-makers:

- why the TGAT proposals were framed in the way they were
- the differences between what was proposed and what was implemented
- the limitations of the current system – and what can be done about them.

A critical contribution to the continuing debate on how learning is measured, for teachers, student teachers, school governors, policy-makers and anyone with an interest in national curriculum assessment.

Dylan Wiliam is Assistant Principal and Professor of Educational Assessment at King's College London.

ASSOCIATION OF TEACHERS
AND LECTURERS
7 NORTHUMBERLAND STREET
LONDON WC2N 5RD

TEL 020 7930 6441
FAX 020 7930 1359
E-MAIL info@atl.org.uk
WEB www.askatl.org.uk