Once you know what they've learned, what do you do next? Designing curriculum and assessment for growth.

Dylan Wiliam

Institute of Education, University of London

Introduction

In this chapter, I describe the development and implementation of a system of age-independent levels of achievement in the national curriculum of England and Wales. The chapter begins by outlining the introduction of the national curriculum, and the brief given to a working group that was to advise the government on the assessment and reporting of student achievement. The next sections discuss some of the alternatives explored by the working group, and in particular, the choice between age-dependent and age-independent levels of achievement. While the idea of age-independent levels of achievement may be unfamiliar, there is substantial research evidence that in many subjects, achievement does appear to be relatively independent of age, and some of this evidence is reviewed briefly. In the final sections, the implementation of the assessment system, and its impact on practice is described.

The National Curriculum of England and Wales

In June 1987, the British Government announced its intention to introduce a national curriculum in England and Wales (Department of Education and Science & Welsh Office, 1987). 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 of age). The foundation subjects were English, mathematics, science, technology, history, geography, art, music and physical education, and in Wales, Welsh. In addition pupils in secondary schools (ages 11 to 16) would also study at least one modern foreign language. The national 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 (pp 9-10).

Alongside the attainment targets, programmes of study would be specified for each foundation subject that 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 (p 10).

The achievement in terms of the attainment targets was 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 of the school.

It was envisaged that much of the assessment at 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 (p 11).

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

It is also clear from the outset that the Government intended that schools would be required to publish, in aggregated form, the results of national curriculum assessments. The reasons given were

> [to] enable schools to be more accountable for the education they offer to their pupils, individually and collectively. The Governing body, headteacher and the teachers of every school will be better able to undertake the essential process of regular evaluation because they will be able to consider their school, taking account of its particular circumstances, against the local and national picture as a whole. […] Parents will be able to judge their children's progress against agreed national targets for attainment and will also be able to judge the effectiveness of their school. LEAs [Local Education Authorities] will be better placed to assess the strengths and weaknesses of the schools they maintain by considering their performances in relation to each other, and to the country at large, taking due account of socio-economic factors … (pp4-5).

Although the June 1987 document was technically, a consultation document, it was clear that the government was unlikely to deviate from its plans, because in September 1987, it established the "Task Group on Assessment and Testing" (TGAT) with a

In September 1987, the Secretary of State for Education and Science and the Secretary of State for Wales commissioned the "Task Group on Assessment and Testing" (TGAT), to provide advice:

> 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 (Department of Education and Science & Welsh Office, 1987 p26)

Specifically, they were asked to advise:

> 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 (National Curriculum Task Group on Assessment and Testing, 1988a appendix A).

The group published its recommendations in January 1988 (National Curriculum Task Group on Assessment and Testing, 1988a) and produced three supplementary reports which were published later in the same year (National Curriculum Task Group on Assessment and Testing, 1988b).

A full description of the work of the Task Group, and its theoretical underpinnings, is beyond the scope of this paper (for an extended discussion, see Wiliam, 1993a). In this paper, I want to focus on a key recommendation of the group—the idea of a system of age-independent levels of achievement—and its significance for the design of curriculum and assessment that focuses on student progression.

Age-dependent vs. age-independent levels of achievement

It seems clear that when the Government issued the 1987 consultation document, it was thinking of the assessments as a series of "benchmarks" (Nuttall, 1989 p. 50): performance standards against which a student could be measured and found either satisfactory or wanting. In the original consultation document (Department of Education and Science & Welsh Office, 1987), attainment targets were defined as establishing "what children should **normally** be expected to know, understand and be able to do at around the ages of 7, 11, 14 and 16" (p9 – my emphasis).

The difficulty with such simple benchmarks is that if they are sufficiently demanding so that they provide real challenges for the most able, then they are so far beyond the reach of most students that the 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 they can attain a 'satisfactory' score with little effort.

This may have been realized 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** to raise expectations, particularly of pupils of middling achievement who frequently are not challenged enough, as well as stretching and stimulating the most able (Department of Education and Science & Welsh Office, 1987 p. 10, emphasis in original).

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 underexpectation by teachers of what their pupils can achieve" (p. 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 underexpectation, but it seems likely that such a system would be less demotivating than a single benchmark for each age group.

The group considered two main categories of approach. The first was that the reporting structure should consist of *independent* reporting scales at each of the reporting ages, while the second involved a framework where the reported scores at any one key stage are directly related to those at other key stages.

Age-specific scales

Systems of age-specific scales are, of course, very familiar all over the world. The traditional literal grades used for school assessment (e.g., A, B, C, D, and F) are age- or grade-specific in that a grade "B" in sixth-grade is meaningful only in the context of sixth-grade. Whether such a grade equates in any sense to a grade "A" in fifth-grade or a grade "C" in seventh-grade is never addressed, and indeed, given the way that curricula are designed, may not even be a meaningful question.

The No Child Left Behind Act of 2001 is also based on the principle of age-specific grades in that it requires that students in grades 3 through 8 are assigned to one of at least three levels (often characterized as below basic, basic, and proficient) although most states appear to have established systems with four distinct levels of achievement (the additional level typically being characterized as "advanced"). The familiarity of such schemes is, of course a tremendous advantage, but they have many substantial difficulties. Here I will confine myself to discussing two: the impact on students, and the interpretability of results.

Impact on students: The ostensible aim of the No Child Left Behind Act, like the Education Reform Act in England and Wales, is to raise student achievement. Such legislative reforms are inevitably highly complex, but a common theme in their justification is the idea that students will be motivated to raise their attainment in order to achieve one of the higher levels or grades. Whether this does in fact, happen, is of course an empirical question, but much recent work in psychology suggests that this may not be the case. In an

extensive research program extending over a quarter of a century, Carol Dweck (2000) has explored how students make sense of their successes and failures in school, and has shown that a crucial characteristic of students who are successful is that they believe ability to be incremental rather than fixed (in other words, they believe that smart is something you get, not something you are). When the assessment system uses the same labels (e.g., below basic, basic, proficient, and advanced), then the students who receive the same label on multiple occasions are likely to think of the label as describing something stable or long-lasting, rather than simply describing the current level of achievement. How serious an issue this is likely to be depends, of course, on how many students will, in fact, be placed in the same category for a number of years. Estimating this, however, is fraught with difficulties, because the answer depends on both factors related to the individual student, and the quality of instruction that students receive, which in turn is confounded with factors of ethnicity and socio-economic status which are likely to interact strongly with individual student motivation in ways that are poorly understood. Nevertheless, the data that did exist at the time (Hart, 1981; Wiliam, 1993a) suggested that the proportion of students placed in the same category for four years was well over 50%.

Interpretability of results: When Kingsbury, Olson, Cronin, Hauser and Houser (2003) compared the performance of students on state tests with vertically scaled reference tests in reading and mathematics aligned to that state's content standards, they found significant internal inconsistencies in many states. For example, in Arizona, the proficiency standard set by the state for mathematics in the third grade equated to the 46th percentile of achievement on the reference test but the standard for eighth grade was equivalent to the 75th percentile on the reference test. So, students at (say) the 50th percentile of achievement would be regarded as proficient in the third grade, but by the eighth grade, students at the same class rank would be regarded as well below proficiency, despite making "normal" progress. Of course, as Kingsbury *et al*. note, there is no way to determine the location of the standard for each grade but such analyses can indicate where the standards are internally inconsistent. This is an important policy issue, because the failure to scale tests vertically is likely to lead to in appropriate policy-making. In Arizona, for example, a reasonable conclusion from the data would appear to be that middle schools were less effective than elementary schools, even though such a conclusion is not warranted by the data.

Age-independent scales

Given the problems with age-specific scales outlined above, it is not surprising that the Task Group looked at alternatives. The most

In mathematics education, it had been widely accepted that, for certain aspects of the subject at least, achievement was not closely tied to age. Two reports from the Assessment of Performance Unit (roughly similar in purpose to the National Assessment of Educational Progress in the USA) in 1980 had shown that high-achieving 7-year-old students out-performed some 14-year-olds on basic arithmetic. One item in particular:

$$6099 + 1 = ?$$

gained some notoriety when it was found that there were many 14-year-olds who thought the answer was 7000, while many 7-year-olds knew the answer to be 6100 (Foxman, Cresswell, Ward, Badger, Tuson & Bloomfield, 1980; Foxman, Martini, Tuson &

Cresswell, 1980). It was this item that led to the idea that there was a “seven-year-gap” between the lowest and highest achieving students in a middle-school mathematics class (Committee of Inquiry into the Teaching of Mathematics in Schools, 1982).

In fact, general competences in mathematics also showed the same, or even greater, variability that had been found by the Assessment of Performance Unit (APU). The Concepts in Secondary Mathematics and Science (CSMS) project had identified a series of 6 levels of understanding of decimals, and in a nationally representative sample found that the variability within each age cohort was much greater than the differences between cohorts (Hart, 1981). In particular, the proportion of students achieving a particular level increased by only 5-10% per year (see Figure 1).

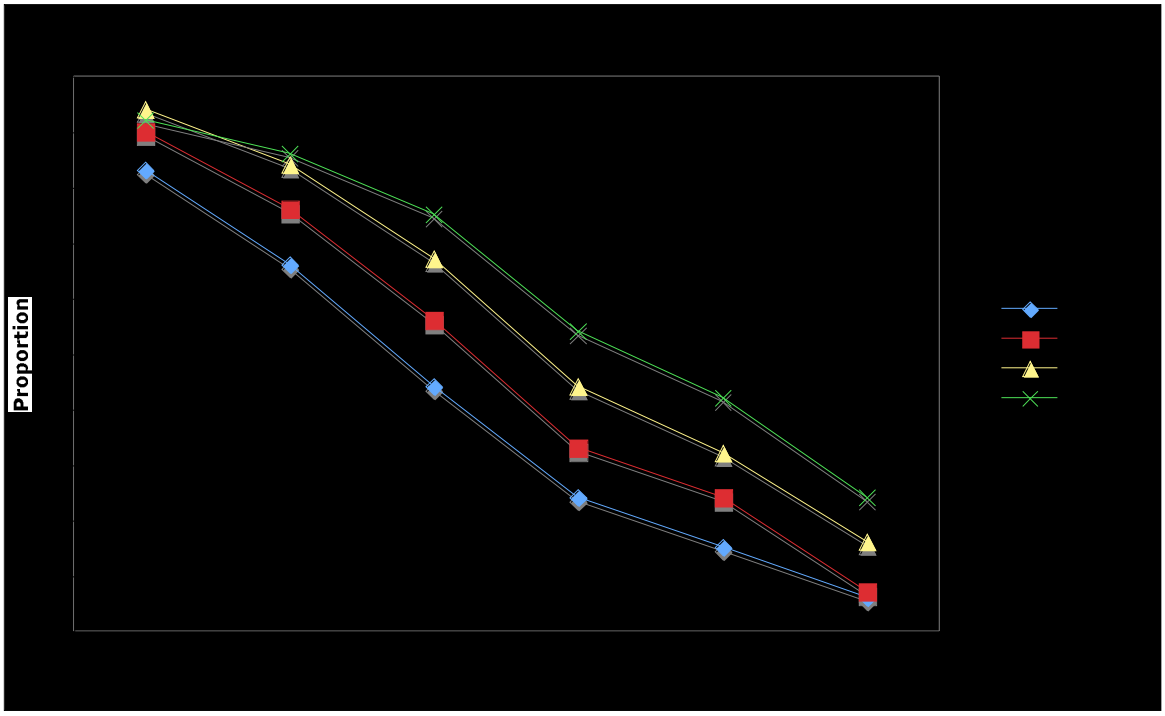


Figure 1: Achievement in Decimals by age found in CSMS (Hart, 1981)

It is also worth noting that such findings are not confined to the United Kingdom. In the mid-1950s the Cooperative Test Division at the Educational Testing Service produced a series of Sequential Tests of Educational Progress (STEP) in reading, writing, listening, social studies, mathematics and science (Educational Testing Service Cooperative Test Division, 1957). The tests were aimed at students from 5th grade to the first two years of college, and were vertically scaled, permitting comparisons to be made across years. The annual increase in achievement in the STEP tests, measured in standard deviations, is shown in Figure 2. Apart from the earliest and latest grades, the typical annual increase in achievement is between 0.3 and 0.4 standard deviations, suggesting that a student at the 95th percentile is as much as ten years ahead of a student at the 5th percentile.

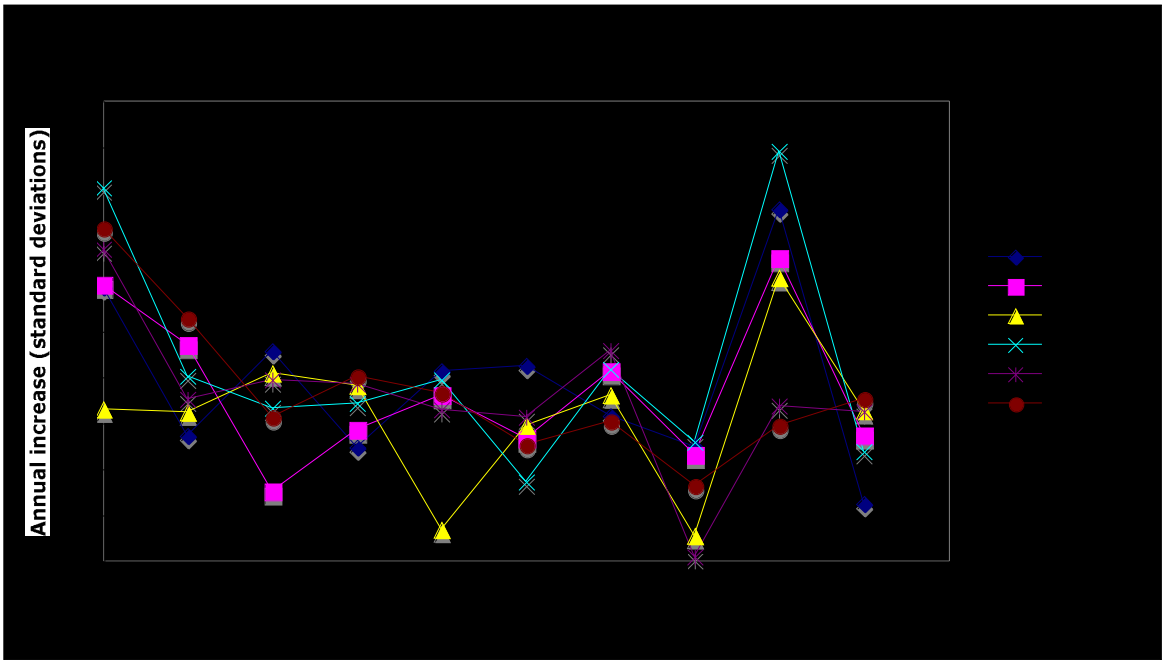


Figure 2: Annual growth in school attainment in the ETS STEP tests (1957)

Other tests show similar properties. Petersen, Kolen and Hoover (1989) discuss the results of scaling the results from the Iowa test of basic skills (ITBS) language usage test (Hieronymous & Lindquist, 1974) for different cohorts of students. For example, a median grade 3 student attains a grade equivalent of 3.5 half-way through the year. However, some of the higher attaining students will have achieved a higher level. The data from the ITBS scaling studies indicate that about 30% of students will, by half way through grade 3, have achieved an attainment equivalent to that achieved by the median fourth-grader at the same time. In a very real sense, therefore, these 30% of students are at least one year ahead of the median. Collecting similar data points for third graders and joining in them up gives us a "grade characteristic curve" for third grade. A similar analysis applied to students in other grades produces a series of such curves (see Peterson *et al*., p. 234). So, for example, in the ITBS language usage tests, the standard associated with average students half way through fourth grade is also just attained by the lowest attaining 5% of students in eighth grade, the lowest-attaining 10% of those in seventh grade, the lowest-attaining 18% of those in sixth grade, and the lowest-attaining 30% of those in fifth grade. On the other hand, the same standard is reached by the highest-attaining 30% in third grade as noted above, and probably by some students in second grade, although this is not recorded. To this extent, what is being measured in the ITBS language usage test would seem to be relatively age-independent. While Petersen *et al.* advocate caution in making such interpretations, it is clearly the case that in some sense, even in language usage, attainment is only loosely related to age. In the ITBS test, one year's growth ranges from around 0.5 standard deviations in third-grade, to around 0.35 standard deviations in 8th grade, which are quite similar to the data for the STEP tests shown in Figure 2. More recent data also confirms that annual growth, when measured in standard deviations, typically ranges from around 0.25 to 0.4. Rodriguez (2004) found that one year's progress in middle-school mathematics on the tests used in TIMSS (Trends in Mathematics and Science Study) was equivalent to 0.36 standard deviations, while the average increase in achievement in mathematics between fourth-grade and eighth-grade on the assessments used in the National Assessment

of Educational Progress (NAEP) was approximately one standard deviation (NAEP, 2006), suggesting that for the NAEP tests, one year's growth is only about one-quarter of a standard deviation.

One way of putting these data into sharp relief is by considering them in terms of what they might say about the instructional sensitivity of such tests. At one end of the scale, we can imagine an idealized IQ test in which scores do not increase with age. Such a test would be completely insensitive to instruction. At the other end of the scale, we can imagine a test in which the lowest-achieving students gain higher scores than the highest scoring students in the class achieved on the same test a year earlier. Such a situation might, for example, arise in a test which assessed material taught only in eighth-grade, and which students would not learn outside school. With such a test, the lowest-performing eighth-graders would outscore the highest-performing seventh graders. For such a maximally sensitive test, it seems plausible, for the sake of simplicity, that the eighth-grade mean would be four standard deviations higher than the seventh-grade mean. If we assign a "sensitivity to instruction" rating of zero to the completely insensitive test, and assign a rating of 100 to the maximally sensitive test, we can construct an index of sensitivity to instruction, as shown in Table 1. Of course these are very "rough and ready" calculations, and, as we will see below, the standard deviation increases with age, so that the index is not independent of age. Nevertheless, the values of the index of instructional sensitivity in Table 1 do show that typical tests cluster towards the lower end of the scale

Test	Sensitivity to instruction score
Completely insensitive test	0
NAEP	6
TIMSS	8
ETS "STEP" tests	8
ITBS	10
Maximally sensitive test	100

Table 1: Values of instructional sensitivity index for selected tests

Age-independent levels in curricula

One of the earliest curriculum designs that explicitly recognized the variability of achievement within an age-cohort was the 'Dalton Plan', which had been developed by Helen Parkhurst at Dalton High School in Massachusetts in 1916. The main feature of the Dalton Plan was that the entire curriculum was broken up into units that might take as much as one month to complete. A student would be given the unit of work in the form of 'self-teaching' assignments. In order to cater for differences between students, the work was sometimes differentiated, as the following extract from Helen Parkhurst's account of the development of the plan illustrates:

> In cases where experience has revealed a marked disparity of intelligence between the pupils of the same age and form, it is sometimes well to modify the assignment in order to bring it within the reach of, say, three different categories. The minimum assignment will merely require the essentials for a [firm] foundation, and its execution should not put too great a strain upon the least gifted pupils in the class. The medium assignment would be given to the next group of moderately intelligent children, while the maximum assignment would be reserved for star pupils. As any

> individual gains ground or develops intellectually, which is a common phenomenon after the Dalton Plan has been in operation for some time, he can be moved from the minimum to the maximum group. But it should never be forgotten that uniformity is not at all synonymous with progress. (Parkhurst, 1922; p. 48)

The Kent Mathematics Project (KMP) explicitly built on the key ideas of the Dalton Plan (Banks, 1985 p58) and began in Ridgewaye School in Southborough, Tunbridge Wells, Kent in 1966 and was adopted more widely through Kent from 1970 onwards (Pennycuick & Murphy, 1988, p51). It was designed to offer "personal mathematics courses for all students between 9 and 16 years which are separately extracted from a material-bank of programmed booklets, tapes and worksheets organised into mathematics levels and areas" (Banks, 1975, p2). Each task was given a level from 1 to 9, with the levels intended to relate to the average "conceptual development" of a student of average ability. Level 1 was intended for average 9-year-olds, with each subsequent level being appropriate for students one year older. While a particular level was defined in terms of the ability of the average student at a particular age, it was also *explicitly* assumed that such materials would be appropriate for older, but lower-attaining students as well as younger, high-attaining students. A theoretical model of the relationship between ability, chronological age and KMP level, developed by KMP is shown in figure 3 (Banks, 1980). An average student would therefore be expected to start the first year of secondary schooling working on level 3 material, and would progress through to level 6 by age 16, while a low attainer would begin with level 1 and reach level 4. According to the model, a very slow learning 11-year old would find level 1 too demanding, and in response to this, some new special material, less demanding than level 1 was developed.

In this sense, the KMP levels framework is consistent with a model of progression that assumes that students with less ability learn more slowly. Students who are above average in ability attain one level each year. Average students achieve six levels in seven years while slow learners achieve about three levels in five years. In KMP, therefore, an activity at level 3 is regarded as appropriate for a gifted 10 year-old, an above average 11 year-old, an average 12-year-old, a below average 13-year-old, a slow 14-year-old, and a very slow 16-year-old. In this sense, KMP clearly forms one of the first examples of a well-articulated and explicit educational assessment framework of age-independent levels of attainment.

Figure 3: Relationship between ability, chronological age and KMP levels

These ideas were developed further within the Graded Assessment in Mathematics (GAIM) project, which sought to develop a national assessment system for students aged 11 to 16. Early on in the work of the GAIM development team, two "ground rules:"

The highest levels of the assessment scheme should be equivalent to the grades of the national school-leaving examination taken by students at the age of 16

The system should be designed so that all students had a reasonable chance of achieving one level per year.

Originally, it was hoped that somewhere between 9 and 13 levels might suffice, but after taking account of the data from the CSMS project about the rate of improvement of skills over time, a model of 15 levels was adopted. This model is encapsulated in Figure 4, which shows the percentage of the cohort able to achieve each of the 15 levels (the horizontal scale represents the years of secondary school in England and Wales, with year 1 corresponding to sixth-grade, year 2 to seventh-grade, and so on).

As can be seen, the GAIM model does allow the majority of students a reasonable chance of achieving one level a year. However, there are places where requirements for the spacing of the levels required by the equivalence to the grades of the school-leaving examinations taken by the end of the fifth-year of secondary schooling result in the gap between adjacent levels being two years in some places. This could, of course, be addressed by having a greater number of levels, but defining a greater number of levels would have posed significant difficulties in defining the levels adequately. Setting the number of levels at 15 was therefore a compromise, but one that seemed reasonable in the circumstances.

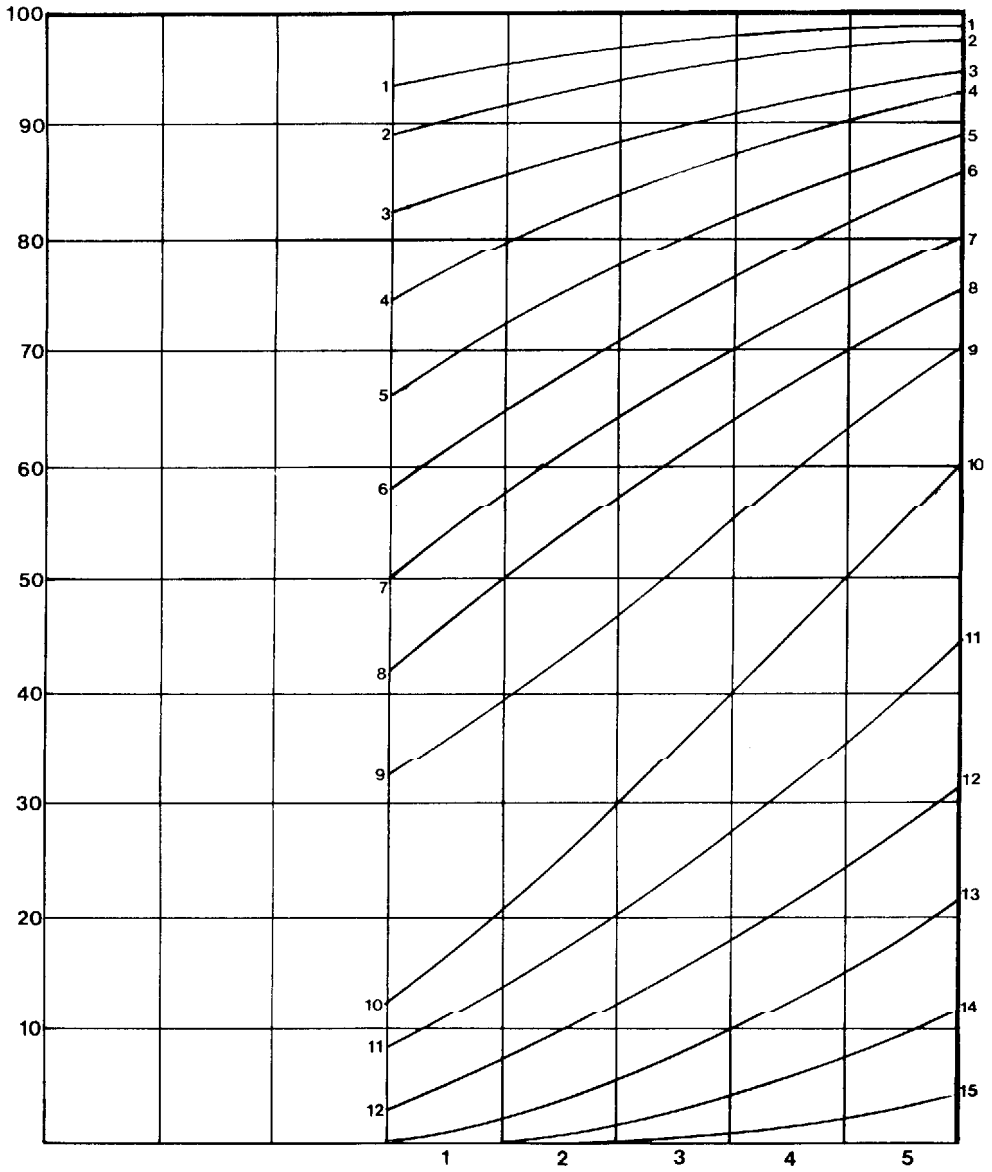


Figure 4: Design of assessment scheme for GAIM project (see text)

The GAIM model was being implemented in schools in the fall of 1987 when the Task Group was formulating its advice to the government. At one of the Task Group's meetings, Margaret Brown, director of the GAIM project, gave evidence about graded assessment schemes in general and GAIM in particular. At one point, she was asked how many additional levels would be required to cover the elementary school grades in the GAIM system. 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 (Wiliam, 2001).

While distinguishing meaningfully between 20 levels of achievement might have been possible with mathematics and science, it was not at all clear that this was possible with 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 10-level model is encapsulated in Figure 5 below.

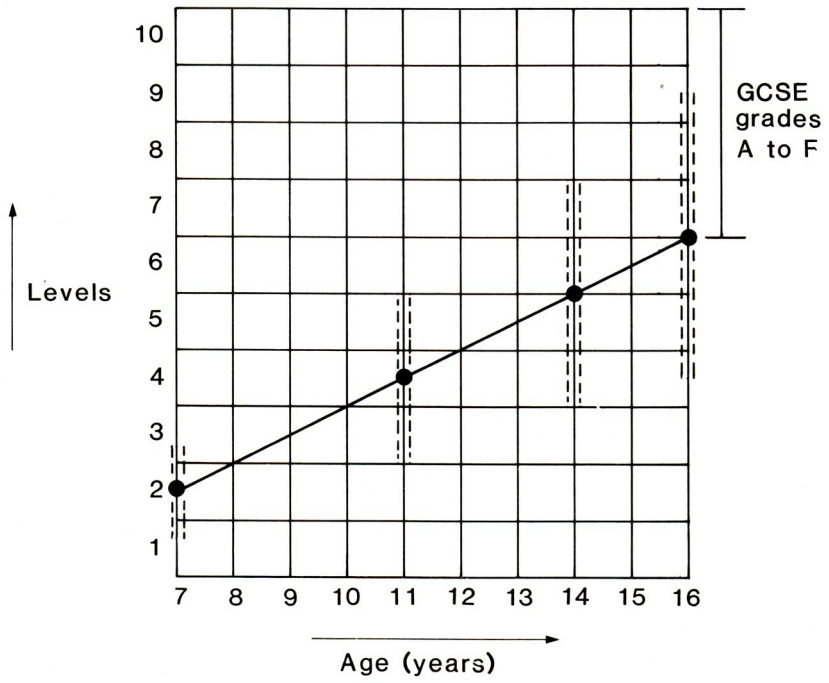
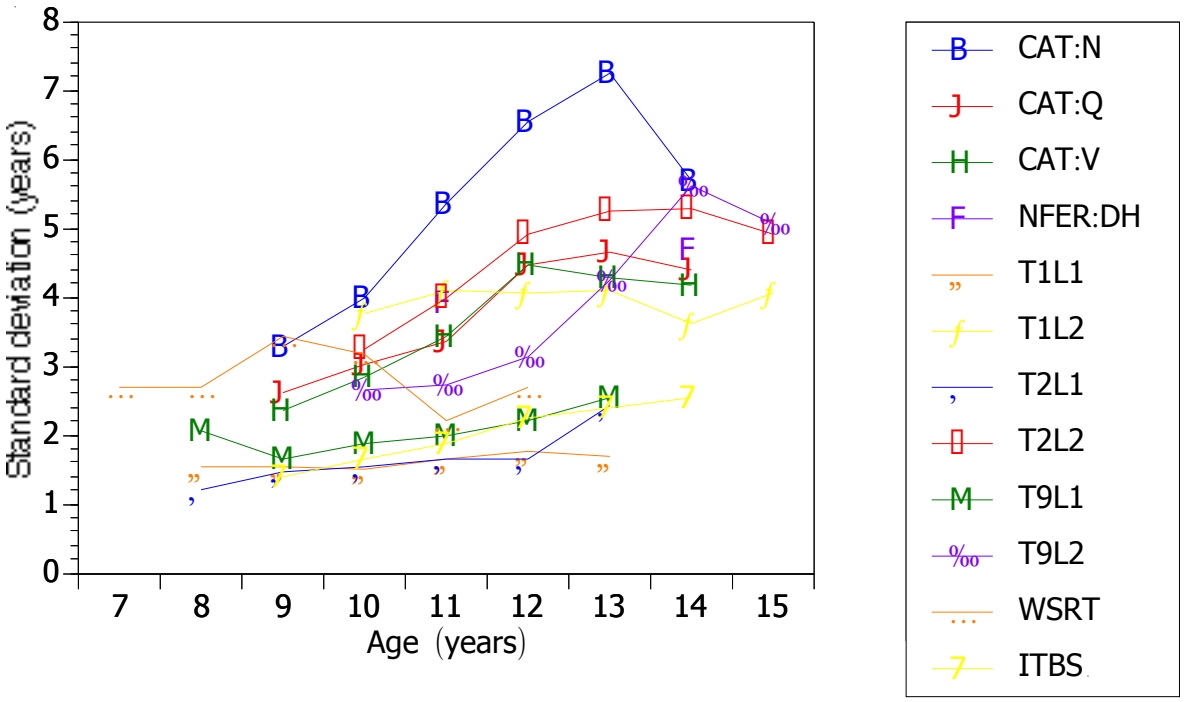


Figure 5: Assessment framework proposed by the Task Group on Assessment and Testing

In the assessment framework proposed by TGAT, most 7-year-olds would achieve level 2, although some would achieve level 3 and others would achieve only level 1. At age 11—the end of elementary schooling in most districts in England and Wales—most students would achieve level 4, although the achievement of students would range from level 2 to level 6. At age 14, the range would be levels 3 to 7, and at 16, the range would be levels 4 to 10. While setting the standard for each level was an arbitrary matter, the Task Group was quite aware that once this had been done, the proportion of students achieving each of these levels at different ages was not arbitrary. For example, we can set a standard so that a particular proportion of 11 year olds achieve this standard. Once we have done this, however, the proportion of 7- or 14-year-olds who can achieve this standard is an empirical matter. In publishing its proposals, the Task Group speculated that the range of levels achieved by 80% of the population would be as indicated by the dotted lines in Figure 6.

The implicit assumption in the TGAT framework is that the spread of achievement in the cohort increases as the cohort ages—a fact that was first observed for physical measurements such as height over a century and a half ago by Quetelet (1835). Cognitive measurements indexed by time (e.g. by grade or age) show much the same picture (see for example, Williamson, Applebaum, & Epanchin, 1991) although some order-preserving measures (e.g., using item-response modeling) do not (Yen, 1986). Since the TGAT proposal was implicitly indexed by age (the levels are a linear transformation of age) the proposal seems plausible, although to ascertain the feasibility of the framework it is necessary to look at data from other sources.

The largest easily accessible body of data about the attainment of whole cohorts of school-students is to be found in the pages of norms provided with commercially available standardized tests. Although these norms have been derived through sophisticated processes involving a number of assumptions that can make interpreting them difficult, such tables do represent claims about the relative performance of students of different ages. For example, a raw score of 38 might be average for an eight-year-old in a particular test, and so would be converted to a standardized score of 100 for age 8. The mean raw score for ten year-olds might be 46, but this same raw score is likely to be obtained by some eight-year-olds. If we find that a raw score of 46 at age 8 is associated with a standardized score of 115, then we can infer that 8-year-old students who are one standard deviation above the mean, are, in a very real sense, two years ahead of average for this test. We can then look at the standardized score for age 8 that is given to average raw scores at other ages in the tables. The slope of the least-squares line of best fit for these points then gives an estimate of the spread of attainment within the cohort, measured in years. The results of this analysis applied to a variety of standardized tests used widely in the UK are presented graphically in Figure 6 (see Wiliam, 1992a for further details). A similar analysis for the data from ITBS that was discussed above is included in Figure 6 for reference.



Key:

CAT:N	Cognitive abilities test non-verbal battery (Thorndike et al, 1986)
CAT:Q	Cognitive abilities test quantitative battery (Thorndike et al, 1986)
CAT:V	Cognitive abilities test verbal battery (Thorndike et al, 1986)
CSMS(M)	Concepts in Secondary mathematics and Science (Hart, 1980)
NFER:DH	NFER non-verbal test DH (Calvert, 1958)
SRT (A)	Suffolk reading test form A (Hagley, 1987)
SRT (B)	Suffolk reading test form B (Hagley, 1987)
T1L1	Profile of mathematical skills test 1 level 1 (France, 1979)
T1L2	Profile of mathematical skills test 1 level 2 (France, 1979)
T2L1	Profile of mathematical skills test 2 level 1 (France, 1979)
T2L2	Profile of mathematical skills test 2 level 2 (France, 1979)
T9L1	Profile of mathematical skills test 9 level 1 (France, 1979)
T9L2	Profile of mathematical skills test 9 level 2 (France, 1979)
TGAT	Task Group report (National Curriculum Task Group on Assessment and Testing, 1987)
WSRT	Wide-span reading test (Brimer, 1983)

Figure 6: Increase of spread of achievement with age for a range of standardized tests

The data summarized in Figure 6 support the idea that the spread of achievement within the cohort increases with age. For some tests, such as the ITBS and some of the Profile of Mathematical Skills series (France, 1979), the rate of increase is quite small, while for others, notably the three batteries within the Cognitive Abilities Test (Thorndike *et al*., 1986), the increase is quite rapid. However, to a first approximation, the growth of spread, as measured by standard deviation, is reasonably linear, the major exception being the Wide-Span Reading Test (Brimer, 1983), in which the spread of achievement appears to be reasonably constant across age. This suggests that a simple first-order model is provided by assuming that achievement age is normally distributed about chronological age, with a standard deviation proportional to the chronological age. For the ITBS, the constant of proportionality is around one-sixth, while for the Cognitive Abilities Test, the constant would be at least one-third. The model proposed by TGAT is between these two extremes, with a constant of proportionality around one-fourth. In order to visualize these outcomes, Figures 7, 8 and 9 display graphically the distribution of attainment resulting from various values of the constant of proportionality. In Figure 7, the value of the constant is one-tenth. In Figure 8, the standard deviation is assumed to be one-fifth of chronological age, and in Figure 9, it is assumed to be one-fourth the chronological age.

If one asked teachers, and indeed, psychometricians, which of the three figures represents the distribution of achievement on tests in widespread use it seems likely that most would choose Figure 7, and would probably resist the idea that the distribution was anything like that in Figure 8, let alone Figure 9. In fact, as has been shown above, the constant of proportionality in Figure 9—i.e., one-fourth—is at the lower end of the range suggested by the data presented in Figure 6. For most tests in widespread use, therefore, the distribution of achievement will be even flatter, and less differentiated than in Figure 9.

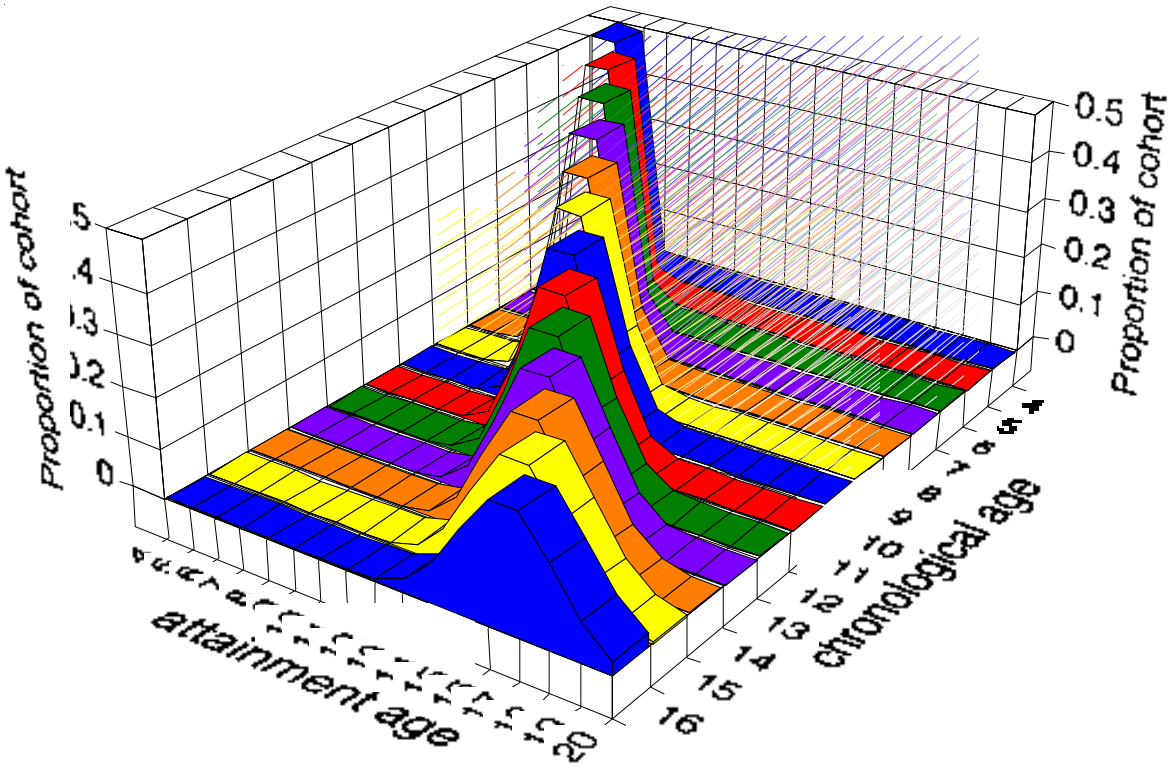


Figure 7: Distribution of attainment for SD at one tenth of chronological age

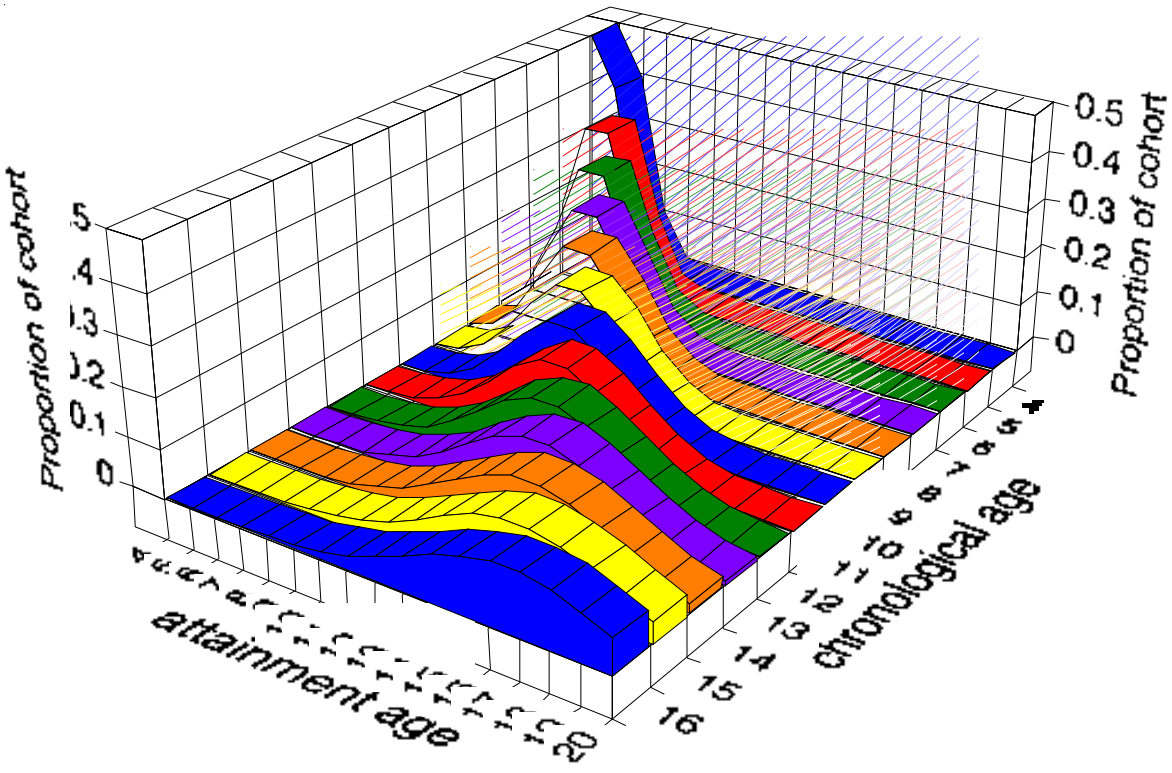


Figure 8: Distribution of attainment for SD at one fifth of chronological age

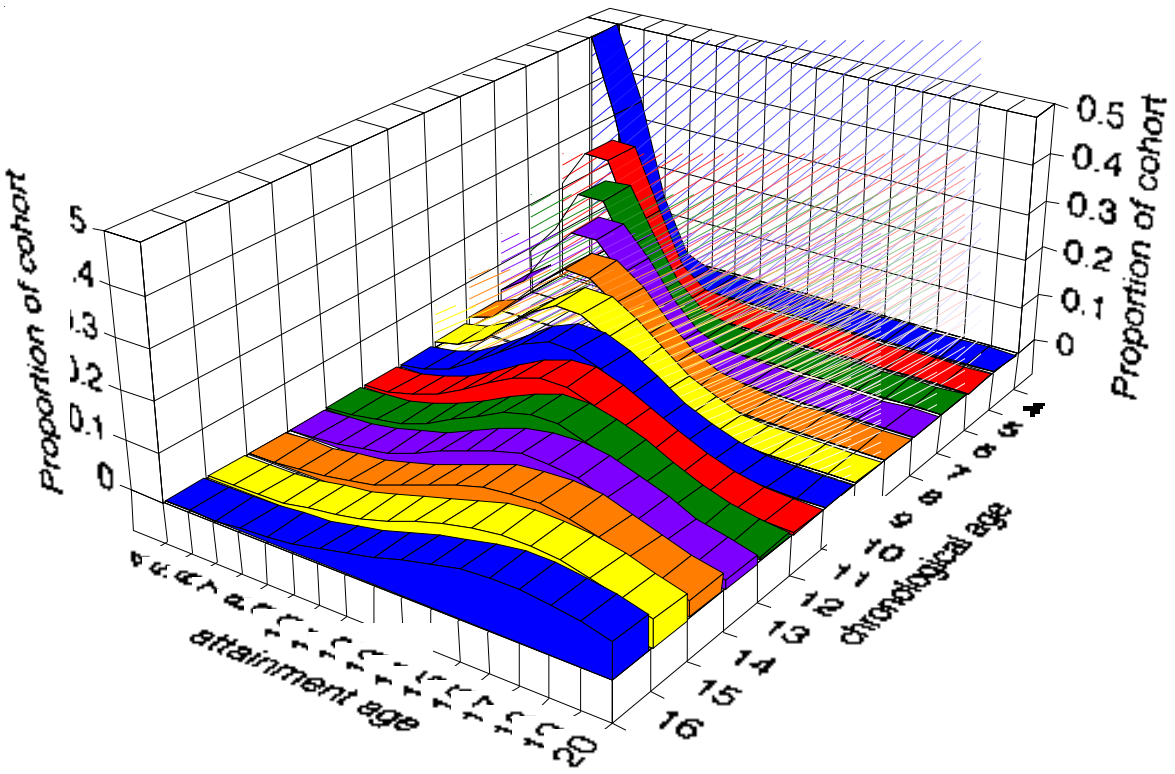


Figure 9: Distribution of attainment for SD at one fourth of chronological age

Implications for the design of curriculum and assessment

Although recommendations of the Task Group were not formally accepted until 7 June 1988 (Secretary of State for Education and Science, 1988), working groups had been developing the national curricula for mathematics and science along the ten-level model since late fall 1987. Why work on developing the national curricula for these two subjects started ahead of those for other subjects is not clear, but the relative ease of operationalizing a framework of age-independent levels of achievement in mathematics and science probably played a part in the subsequent national acceptance of the scheme. It was probably also important that the "rough speculation" of the Task Group about the spread of achievement within the cohort was not unreasonable, and that the director of the Graded Assessment in Mathematics Project, Margaret Brown, was a member of the mathematics working group. While these contingent issues were probably helpful, some of the structural features of the ten-level framework were also important.

In traditional approaches to curriculum specification, the tendency is to define curriculum "horizontally," by looking at what could "go into," say, the sixth grade science curriculum. By focusing on the "vertical" dimension, the TGAT framework encouraged the developers to look at strands of development over many years. While this approach did not make it impossible for curriculum developers to simply re-package the existing curriculum in terms of the ten-level framework, it did create a strong pressure to think of the curriculum in developmental terms.

Of course, while the requirement to think of the curriculum in developmental terms does provide some structure, it is far from a strait-jacket. While some work on concept hierarchies had been done in both mathematics and science (Hart, 1981; Shayer & Adey, 1981), it was clear that these hierarchies were much more the result of particular choices about curriculum sequencing than as a result of something inherent in the subject, or related to levels of psychological processes. For example, Denvir and Brown (1986) found that there were virtually no students who could subtract one number from a larger number without being able to count backwards by one. In this sense, the hierarchy formed by these two skills would appear to be relatively independent of curricular experience. However, other kinds of observed hierarchies are the result of curricular choices, and other curricular choices, as coherent in terms of the nature of the discipline, or underlying psychological processes, would be possible. As an example, consider the traditional sequencing of multiplication and division of small whole numbers. In curricula world-wide, multiplication is taught before division, on the grounds that multiplication is a pre-requisite skill for division calculations. However, the CSMS project (Hart, 1981) found that conceptually, division was easier than multiplication. One of the instruments the CSMS team used to probe the understanding of children's number concepts was to give them calculations, and ask the students to construct stories based on the calculation. They found that many students could construct stories for calculations such as $12 \div 4 = 3$. Typical examples here were, "There were 12 sweets shared between 4 people so they got 3 each." However, they found that some of these students could not construct plausible stories for the equivalent multiplication calculations. Given $3 \times 4 = 12$, students responded with stories such as "Jane had 3 sweets and John had 4 sweets and they timesed them to get 12 sweets." So while computationally, multiplication appears to precede division, conceptually, it would appear to be the other way round. The appropriate sequencing of multiplication and division would therefore depend on whether the computational or conceptual aspects were emphasized.

A similar relationship exists between differentiation and integration in calculus. In almost all teaching of calculus, differentiation is taught before integration, because it is computationally easier. However, conceptually, the area under a curve is much more accessible than the idea of the gradient of a curve at a point. In the current approach to calculus, differentiation must precede integration, but there are valid approaches to teaching calculus where integration might be introduced before differentiation.

Any developmental progression, therefore, is much more a reflection of historical contingencies than anything inherent in the discipline, or of the underlying processes. The task for the working groups developing the curricula for each subject was therefore to integrate knowledge about what was known about the relative difficulty of different aspects of the subject, in the light of the likely curriculum exposure that influenced the difficulty. Ideally, it was also important to ensure that the resulting sequence was useful in providing information to teachers about important aspects of students' development

The final version of the science national curriculum proposed the following statements of attainment for a strand on the nature of light (Black, 1995):

1) Know that light comes from different sources

2) Know that light passes through some materials and not others, and that when it does not, shadows may be formed

3) Know that light can be made to change direction, and that shiny surfaces can form images

4) Know that light travels in straight lines, and this can be used to explain the formation of shadows

5) Understand how light is reflected

6) Understand how prisms and lenses refract and disperse light

7) Be able to describe how simple optical devices work

8) Understand refraction as an effect of differences of velocities in different media

9) [nothing new at this level]

10) Understand the processes of dispersion, interference, diffraction and polarisation of light

The developmental strand shown here for light has several important features. First, and most obviously, it focuses attention on the notion of *progression*. While it is certainly possible to overstate the impact of this shift of attention, the emphasis on progression rather than mastery tends to change the orientation of both student and teacher away from mastery (did the student master the material or not) and towards locating the student somewhere on a continuum. The second important feature of the strand is that the strand takes into account the psychological processes involved in learning about light. While there is no requirement for the strand to be developmentally coherent, any inconsistencies are more immediately apparent. When psychological processes underpin the development of curriculum in this way, then, it is more likely that the developmental sequences are not just *developmentally coherent*, but also *instructionally tractable*. In other words, once we know where a student is, the very nature of the continuum helps us determine what kinds of instructional experiences should follow (Wiliam, 1993b).

A third feature of the strand is that it is not particularly dense in terms of students' development. Not all the things that need to happen in students' learning about light are here; only the most important and significant indices of the stage the student has reached are included. Indeed, because it was felt that, at level 9, there was nothing important enough to be worth checking on, nothing appeared. The statements within the strand are therefore more like "checkpoints" on an orienteering course that a developmental progression that all students are expected to follow. This avoids what Lorrie Shepard has called the "thousand mini-lesson problem" (Shepard, 2007). Where assessment reveals that students have failed to display mastery of a significant proportion of a domain, without some coherent map of development, all the teacher can do is attempt to rectify each of the problems one-at-a-time.

Of course, this model worked better in some subjects than others. In many school history curricula, it was typical to adopt a broadly chronological approach, so that students studied early cave dwellers, Ancient Egypt and Ancient Rome in elementary school, the Vikings in early middle school, and the Victorian era at the end of middle school. However, this simplistic model of chronological accumulation was rendered absurd by the ten-level

model, with its notion of age independent levels of achievement. Level 3 was meant to be the level attained by high-achieving 7-year-olds and low-achieving 14-year-olds, but what sense did it make to say that being very good on the Romans was somehow equivalent to being average on the Vikings, and below average on the Victorians. The absurdity of such a model forced the developers of the history curriculum to think about what it is that develops when someone makes progress in history.

When it published its final report, the National Curriculum History Working Group (1990) proposed four attainment targets for history across the 5 to 16 age range:

Understanding history in its setting

Understanding points of view and interpretations in history

Acquiring and evaluating historical information

Organising and communicating the results of historical study.

The detailed "statements of attainment" at each of the ten levels proposed for the first attainment target were as follows:

1) Recognise everyday time conventions.

2) Place a few straightforward events in chronological sequence; demonstrate, by reference to the past, an awareness that actions have consequences.

3) Demonstrate awareness of a variety of changes within a short time span; demonstrate an awareness of human motivation illustrated by reference to events of the past.

4) Employ appropriate chronological conventions by using time-lines or other diagrammatic representation of historical issues; understand that historical events usually have more than one cause or consequence.

5) Demonstrate a clear understanding of change over varied time periods; understand that historical events have different types of causes and consequences.

6) Recognise some of the complexities inherent in the idea of change, when explaining historical issues; when explaining historical issues, place some causes and consequences in a sensible order of importance.

7) When explaining historical issues, show a detailed awareness of the idea of change; when examining historical issues, can draw the distinction between causes, intentions, motives and reasons.

8) Apply extensive understanding of change to complex historical issues; produce a well-argued hierarchy of causes for complex, historical issues.

9) Demonstrate an awareness of the problems inherent in the idea of change; demonstrate an awareness of the problems inherent in the idea of causation.

10) Demonstrate a clear understanding of the complexities of the relationship between cause, consequence and change.

Now of course historians might disagree about the particular choice of attainment targets, and the particular statements of attainment that were chosen to show the nature of development in each of the attainment targets. But I think it is clear that the discipline of being forced to think in terms of developmental strands resulted in a far deeper conceptualization of the nature of historical thinking than is found in most history curricula. In particular, the need to identify developmental strands made it more difficult for the curriculum developers to add content just because it was felt to be important (or at least made it more obvious if they did so). Even in English Language Arts, where the initial resistance to the 10 level scale was most profound (see, for example, Barrs, 1990), a reasonable consensus quickly emerged about the nature of progression. In writing, for example, progression was defined as "a growing ability to construct and convey meaning in written language matching style to audience and purpose" (National Curriculum Council, 1989 p. 45).

As development work on assessment progressed, it was clear that the differences in the models of progression that had been adopted in the different subjects had a substantial impact on the ease of applying the model. In mathematics, while the curriculum was defined in terms of skills, the assessments emphasized students' ability to think mathematically, rather than to recall knowledge, or implement rehearsed routines. High-achieving students were therefore able to answer successfully items on material they had not been taught by using high-level general reasoning skills (indeed, one group of Norwegian mathematics educators who saw the mathematics items were appalled, because the items looked to them like those in IQ tests). As a result the variability of achievement was relatively high. In science, the assessments tended to concentrate more on the ability to recall and use scientific concepts that students had been taught in school, and therefore curriculum exposure appeared to be more important than reasoning ability resulting in the variability of achievement being less than in mathematics. In English Language Arts, the assessments focused more on students' ability to respond to text, and therefore maturity seemed to be more important than either reasoning power or curriculum exposure.

Of course these outcomes are contingent rather than necessary. It would be perfectly straightforward to design mathematics assessments that emphasized curriculum exposure, science assessments that emphasized maturity, or English Language Arts assessments that emphasized reasoning power. Where achievement in a subject is defined in a way that emphasizes reasoning power, student achievement will be highly variable, and where it is defined in terms of maturity, or curriculum exposure, students achievement will be less variable.

The unfamiliarity of the 10-level scale, and teacher unrest related to the workload associated with the national curriculum assessment (see Wiliam, 1995 for a discussion of the causes and effects of this dispute) led the government to commission a review of the curriculum and the assessment framework in April 1993. An interim report was submitted in July 1993, and a final later the same year (Dearing, 1994). The specific issue of thr ten-level scale was raised by only 30% of those who responded to the invitation to contribute to the debate around the national curriculum and its assessment, and these were equally divided between those arguing for the retention of the scale, and those advocating its

replacement with traditional standardized end-of-key-stage tests at ages 7, 11, 14 and 16, where no attempt is made to articulate the relationship between the tests taken at different ages. In the end, the government decided that the scale should not be used for the national assessment of 16 year olds, preferring instead the nine-point grade scale that was already in place removing the need for the highest two levels. However, the government decided that the resulting eight-point scale should be retained—not least because of the lack of evidence that the alternatives were any better—and remains in place to this day.

An evaluation of the impact of the 10- (now 8-) level scale is beyond the scope of this chapter, but the constraints of the model appear to have exerted a helpful discipline on the developers of the national curriculum in England and Wales. In particular, the focus on progression appears to have restricted the proliferation of content standards that have, in other countries, resulted in curricula "a mile wide and an inch deep" (Schmidt, McKnight & Raizen, 1997 p. 62). Also, reports from the rolling program of national inspections of all public schools (see, for example, Her Majesty's Chief Inspector of Schools, 2006) seem to suggest that it has been an important feature of the increases in student achievement since the national curriculum was introduced. The progressive nature of the model enables teachers and students to focus on where students are in their learning, and what the next steps should be, rather than establishing whether the students did, or did not, master an adequate proportion of the curriculum. Finally, the adoption of a single assessment scale for students of different ages has supported the development of a relatively straightforward definition of "value-added" that has not required complex and opaque mathematical procedures (Wiliam, 1992b). The result has been a profound shift in the extent to which administrators are able to track the progress of students in terms of their learning, rather than via proxies such as percentile ranks, which provide little information about existing learning and future steps.

Conclusion

In this chapter, I have presented a rationale for a particular approach to the development of curriculum, and some empirical evidence in its support. The idea of age-independent levels of achievement is unfamiliar, perhaps even outlandish, to most American readers and yet such an approach to assessment has many advantages. Curriculum developers need to think in terms of learning progressions, which brings a coherence to the curriculum that is rarely achieved when curricula are compiled grade-by-grade (see Smith, Wiser, Anderson & Krajcik, 2006 for a contemporary U.S. application to science learning). Such an approach also recognizes a fundamental truth that is often ignored, which is that student learning is more variable, and slower, than is usually assumed. Assessment with respect to learning progressions supports teaching and learning in a far more direct way than traditional approaches, and also strengthens the likelihood that students will come to see ability in a subject as incremental, rather than fixed. Finally, where student outcomes are used as measures of accountability, age-independent levels of achievement support value-added inferences in a straightforward and direct way.

Inevitably, the mode of age-independent levels of achievement that was implemented in England and Wales was far from perfect, and subsequent political interference has tended to make things worse, rather than better (Black, 1997), so that it would be unwise to import such a system wholesale into a different context. However, the idea that curricula and assessment systems should be designed to support learning is gaining ground in the United

States (Popham, Keller, Moulding, Pellegrino & Sandifer, 2005). The experience from England and Wales is that when standards, curricula, and assessments are designed together, rather than serially, then the internal stresses in the system are reduced, and the possibilities for increased student learning are enhanced.

References

Barrs, M. (1990). *Words not numbers: assessment in English*. Sheffield, UK: National Association of Teachers of English.

Black, P. (1995). 1987 to 1995: The struggle to formulate a national curriculum for science in England and Wales. *Studies in Science Education,* **26**, 158-188.

Black, P. (1997). Whatever happened to TGAT? In C. Cullingford (Ed.), *Assessment vs. evaluation* (pp. 24-50). London, UK: Cassell.

Brimer, A. (1983). *Wide-span reading test manual*, Windsor: NFER-Nelson.

Brown, M. (1980). Place value and decimals. In K.M. Hart (Ed.), *Children's understanding of mathematics 11-16.* London: John Murray. Pp 48-65.

Calvert, B. (1958). *Manual of instructions for non-verbal test DH*. Windsor, NFER-Nelson.

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

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

Denvir, B., & Brown, M. L. (1986). Understanding of number concepts in low-attaining 7-9 year olds: part 1. Development of descriptive framework and diagnostic instrument. *Educational Studies in Mathematics,* **17**(1), 15-36.

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

Dweck, C. S. (2000). *Self-theories: their role in motivation, personality and development*. Philadelphia, PA: Psychology Press.

Educational Testing Service Cooperative Test Division. (1957). *Cooperative Sequential Tests of Educational Progress: technical report*. Princeton, NJ: Educational Testing Service.

Foxman, D. D., Cresswell, M. J., Ward, M., Badger, M. E., Tuson, J. A., & Bloomfield, B. A. (1980). *Mathematical development: primary survey report no 1*. London, UK: Her Majesty's Stationery Office.

Foxman, D. D., Martini, R. M., Tuson, J. A., & Cresswell, M. J. (1980). *Mathematical development: secondary survey report no 1*. London, UK: Her Majesty's Stationery Office.

France, N. (1979). Profile of mathematical skills (levels 1 and 2): teacher's book and tables of norms. Windsor, UK: NFER-Nelson.

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

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

Hagley, F. (1987). *Suffolk reading scale teacher's guide*. Windsor: NFER-Nelson.

Hart, K. M. (Ed.). (1981). *Children's understanding of mathematics: 11-16*. London, UK: John Murray.

Her Majesty's Chief Inspector of Schools. (2006). *Annual report 2005-2006*. London, UK: Her Majesty's Stationery Office.

Hieronymous, A. N., & Lindquist, E. F. (1974). *Manual for administrators, supervisors and counselors – levels edition (forms 5 &6): Iowa tests of basic skills*. Boston, MA: Houghton Mifflin.

Kingsbury, G. G., Olson, A., Cronin, J., Hauser, C., & Houser, R. (2003). *The state of state standards: research investigating proficiency levels in fourteen states*. Portland, OR: North West Evaluation Association.

National Assessment of Educational Progress. (2006). *The nation's report card: Mathematics 2005* (Vol. NCES 2006-453). Washington, DC: Institute of Education Sciences.

National Curriculum Council. (1989). *Consultation report: English*. York, UK: National Curriculum Council.

National curriculum history working group. (1990). *Final report*. London, UK: Her Majesty's Stationery Office.

National Curriculum Task Group on Assessment and Testing. (1988). *A report*. London, UK: Department of Education and Science.
National Curriculum Task Group on Assessment and Testing. (1988). *Three supplementary reports*. London: Department of Education and Science.

Nuttall, D. L. (1989). National assessment: complacency or misinterpretation? In D. Lawton (Ed.), *The Educational Reform Act: choice and control* (pp. 44-66). London, UK: Hodder & Stoughton.

Parkhurst, H. (1922). *Education on the Dalton Plan*. London, UK: G. Bell and Sons, Ltd.

Popham, W. J., Keller, T., Moulding, B., Pellegrino, J. W., & Sandifer, P. (2005). Instructionally supportive accountability tests in science: a viable assessment option? *Measurement: Interdisciplinary Research and Perspectives,* **3**(3), 121-179.

Quetelet, L. A. J. (1835). *Sur l'homme et le developpement de ses facultés, essai d'une physique sociale [On man, and the development of his faculties, an essay on social physics]*. London, UK: Bossange & Co.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education,* **17**(1), 1-24.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: an investigation of U.S. science and mathematics education*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Secretary of State for Education and Science. (1988). Assessment and testing: a reply to Mr Key. In *Parliamentary Written Answers, Hansard, 7 June 1988*. London, UK: Her Majesty's Stationery Office.

Shayer, M., & Adey, P. S. (1981). *Towards a science of science teaching: cognitive development and curriculum demand*. London, UK: Heinemann Educational Books.

Shepard, L. A. (2007). Will commercialism enable or destroy formative assessment? In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Simon, B. (1992). The Education Reform Act: causative factors. In P. Broadfoot, W. B. Dockrell, C. V. Gipps, W. Harlen & D. L. Nuttall (Eds.), *Policy issues in national curriculum assessment* (pp. 28-42). Clevedon, UK: Multilingual Matters.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: a proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives,* **4**(1&2), 1-98.

Thorndike, R.L., Hagen, E. & France, N. (1986). *Cognitive abilities test administration manual*. Windsor: NFER-Nelson.

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

Wiliam, D. (1992b). Value-added attacks? Technical issues in publishing national curriculum assessments. *British Educational Research Journal,* **18**(4), 329-341.

Wiliam, D. (1993a). *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.* Unpublished PhD thesis, King's College University of London.

Wiliam, D. (1993b). Once you know what they've learnt, what do you teach next? A defence of the national curriculum ten-level model. *British Journal of Curriculum and Assessment,* **3**(3), 19-23.

Wiliam, D. (1995). The development of national curriculum assessment in England and Wales. In T. Oakland & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 157-185). Boston, MA: Kluwer Academic Publishers.

Wiliam, D. (2001). *Level best? Levels of attainment in national curriculum assessment*. London, UK: Association of Teachers and Lecturers.

Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal analysis of academic achievement. *Journal of Educational Measurement,* **28**(1), 61-76.