The Big New Idea: Single Level Tests

Chris Whetton

Paper presented at the 34th International Association for Educational Assessment Conference, 8 – 12 September 2008, Cambridge, England.

Presenter: Chris Whetton

Director, Research in Assessment and Measurement

National Foundation for Educational Research

The Mere, Upton Park Slough SL1 2DQ

United Kingdom

c.whetton@nfer.ac.uk

The Big New Idea: Single Level Tests

Background

The National Curriculum Assessment arrangements in England now form a relatively mature system. Their genesis was in 1988 when the British government of the time announced the first major reform of the education in England and Wales for over forty years. This was intended to radically alter the nature of compulsory education in those countries. The motivation of the reform was concerns over many aspects of the process of schooling and its outcomes. There was a growing view that despite increases in resources, standards of attainment had not improved since the Second World War. This poor achievement was despite steadily increasing resources and funding being devoted to education. Compared to many other countries there were very wide ranges of attainment, with large numbers of very poorly attaining students. Prior to that time, teachers, schools and local education authorities had determined the curriculum in each locality, leading to large variations in standards. During the 1970s, some pedagogic practices received high levels of publicity and condemnation. In other spheres of government, the ideology of the time had been based on the introduction of market forces in order to raise standards, and this philosophy was applied to education.

The early years of the introduction of the National Curriculum system were turbulent, with an initial overload in the curriculum and in the demands on teachers. This led to a teachers' revolt in 1993 and a consequent slimming down of the system. The review following from this introduced an organisational pattern which began in 1996 and has continued in roughly the same form until the present.

Current National Assessment System

The structure of the curriculum and education system in England is organised into "Key Stages". There are four key stages, which are: five to seven years old (Key Stage 1); eight to 11 years old (Key Stage 2); 12 to 14 years old (Key Stage 3); and 15 to 16 years old (Key Stage 1). For each of the Key Stages there is an end of key stage assessment. The reporting for this is in terms of broad "levels" which run from 1 to 8 and are very loosely criterion-referenced since they relate to the curriculum structure and content and are intended to be defined by these. Level 2 is the expected attainment of a seven-year-old and level 4 of an 11-year-old, so that a level should represent about two years of learning and development.

For Key stage 4, assessment is provided by the school-leaving certification system, with students taking GCSEs (General Certificate of Secondary Education) in a number of subjects. These examinations are provided by independent examination boards and regulated by a government agency. The reporting is not on the level-based scale, but has its own system of grades..

The remainder of the key stage testing is organised centrally by a government agency, the Qualifications and Curriculum Agency (QCA) under direction from the Government.

Taking each Key Stage in turn, for Key Stage 1, there are tests provided to schools by QCA in English and mathematics which are taken at a time which the school chooses. These tests are marked in the school and the results are used to inform teachers' own judgements. These "Teacher Assessment" outcomes are what are reported to parents and centrally collected. This model has operated only since 2005, replacing the previous system, in which test results were reported.

For Key Stage 2, there are formal written tests at the end of the key stage (around May each year) in three subjects: English, mathematics, and science. Each subject has two or three papers: These are externally marked written tests in English (three tests – one in reading, one in spelling and one in writing (comprising two tasks, a longer and a shorter task); mathematics (three tests – one without calculator, one with calculator, and a mental mathematics test); and science (two tests). For English, reading and writing are reported separately and a combined English result is also given. For mathematics and science a single outcome is given.

A feature of the English system is that the tests are used as accountability measures. This is the case for individual primary schools where the proportion of students achieving level 4 at the end of key Stage 2 is published in "league tables" as well as being used by the schools inspectorate to help form judgements on schools. It is also the case on a national basis, where the overall results are used as a measure of governmental success in improving the education system. Hence the tests have a high stakes nature despite not being particularly important to the life chances of the individuals taking them.

For Key Stage 3, there is a wider range of levels to be included, due to the increasing variance in students' attainments, and the English and science tests cover levels 3 to 7, with mathematics covering levels 3 to 8. Hence, unlike Key Stage 2, several tiers of tests are available and teachers must decide the tier to enter each student for. Science has two papers each at two tiers (3 to 6 and 5 to7) and mathematics has two written papers at four tiers (3 to 5; 4 to 6; 5 to 7 and 6 to 8) and a mental mathematics test with two different tiers (3 to 5 and 4 to 8). English, as ever, is different and a single set of tests covers levels 3 to 7. There are papers in writing (a longer and shorter task), reading and a paper on a Shakespeare play. This is not a simple system.

As with the Key Stage 2 tests, the Key stage 3 aggregated results are published for each school and centrally.

The Current Situation

After it stabilised in the mid 1990s, the National Assessment system has been relatively unchanging. However, this is not to say it is without criticism. These criticisms fall into several groups.

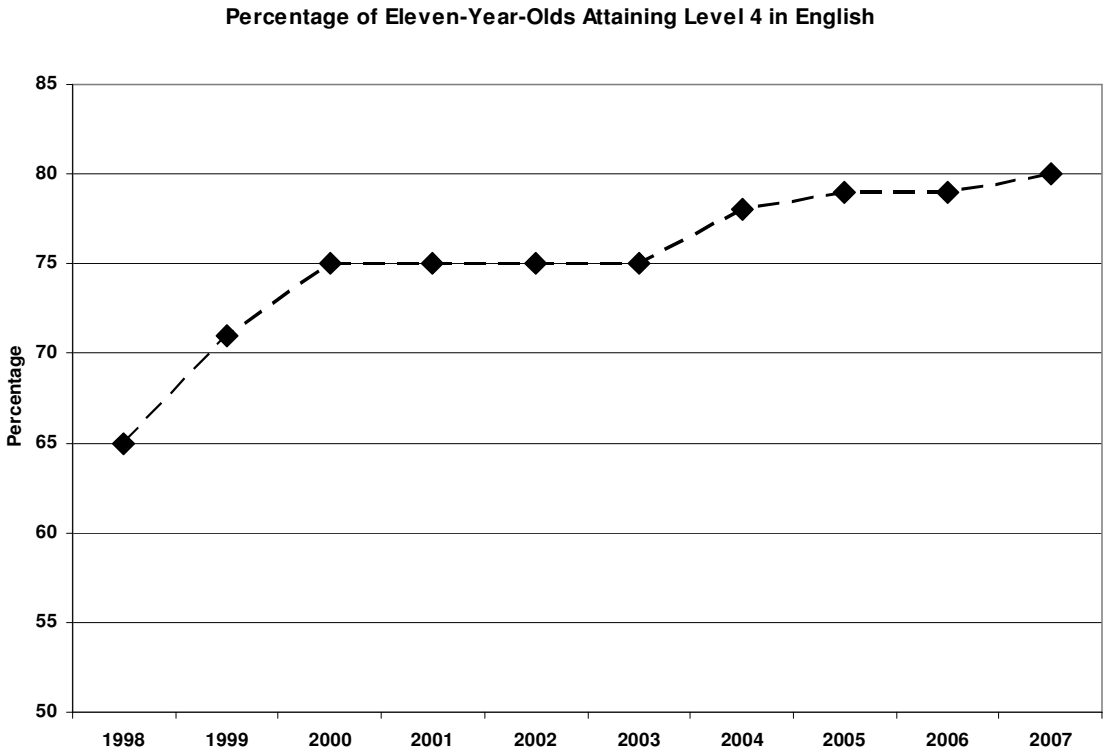
- That the accountability function puts too much pressure on schools
- That the accountability function and the nature of the tests leads to a narrowing of the curriculum
- That the system has developed into having too many purposes for the use of its results so that it cannot adequately serve them all (Newton, 2007, identifies 18 ways in which the test results are being used)
- In particular that the system does not provide a reliable measure of changes in performance over time (Tymms, 2004)
- That the tests put too much pressure on children
- That standards would be raised to a greater degree (and more widely and validly) through an emphasis on "Assessment for Learning"

During 2007 and 2008, these issues were examined in two different arenas: the academic and the political. Under, Robin Alexander, the Primary Review based at Cambridge University examined the condition and future of primary education in England. The Primary Review commissioned academic surveys on all aspects of primary education, including the assessment system. The paper dealing with this concluded:

> "The current system of assessment in England provides information of only low dependability whilst having some negative impacts on teaching and learning. Alternative systems need to be considered. One in which summative assessment is based on teachers' judgements would provide information that is more valid than tests and at least as reliable, but it would be necessary to avoid high stakes being attached to the results by not using them for purposes other than reporting on individual pupils."
>
> "For national monitoring, a regular sample survey, using a large bank of items, would give far more information than is provided by results of individual pupils who have all taken the same test." Harlen (2007)

However, for government, there is a different concern, illustrated in the figure below. This is that the early rapid increases in performance have not been maintained. This can be illustrated by the Key Stage 2 test results for English. These show a rapid increase in the 1990s and then something of a plateau in recent years, or at least less spectacular improvement. A target which should have been achieved in 2002 (80 per cent of pupils at level 4) is still not quite made.

There are arguments that some of the early increase is misleading, (Tymms, 2004; Statistics Commission 2005) but these do not need rehearsing here. From the political perspective the important issue is to reinvigorate the upward movement in performance. If this could also be done in a manner that removed some of the criticism of the existing system, then that would be a bonus.



Making Good Progress

The proposed solution was announced by Alan Johnson, the Secretary of State for Education, in January 2007. This was known as the "Making Good Progress" pilot and within this an assessment system based on teacher assessment and "single level tests". As ever, this was a political decision taken in haste without consultation or any detailed planning. Rather the broad principles were announced and the problem handed over to officials and statutory bodies to provide a workable system.

Making Good Progress (DfES, 2007) stated that the government was interested in:

> "exploring the impact of enabling teachers to enter a pupil for an externally-marked test as soon as they are confident (through their own systematic assessments) that the pupil has progressed to the next level." The tests would be offered to schools at two points in the year and pupils could be entered individually for a test which marks success at one level, and stimulates progress towards the next level."

The tests were to be no more burdensome than the current end-of-key stage "multi-level" tests. Further,

> "the single-level tests would generate the data on achievement that is so important for school accountability. The system would be a one-way ratchet: once a pupil has passed a level, they will never go back, only forward. The model could be a powerful driver for progression, raising expectations for all pupils, motivating them, bringing a sharp focus on 'next steps' and perhaps especially benefiting those who start the key stage with lower attainment than their peers, or

are currently making too little progress. Ultimately, these tests might replace end of key stage arrangements."

Making Good Progress is now a pilot development programme and remains underway. It was originally for two years, but this has been extended and it is expected that it will eventually be implemented to all schools. Currently, ten local authorities and their schools are involved. The pilot has several aspects involving teaching methods, booster classes, reward mechanisms and accountability measures based on progress rather than absolute achievements, but this paper will concentrate on the assessment element, the Single Level Tests.

The first run of the tests was in December 2007. These tests were very quickly constructed, using some existing material, and were not developed through a full defensible process, simply due to the time scales between the announcement and the implementation. Some aspects of the outcomes are public, but there are not yet any evaluation documents published. The subsequent tests in June 2008 did have a proper development process, but at the time of writing their results are not known.

Elements of *Making Good Progress*

This statement provided a very rough outline for a new assessment system. There are within it several features which will require careful conceptual thought and a sound research and development process if the political objectives are in any way to be met.

Within the ***Making Good Progress*** document, it is proposed that there should be a new type of tests and a new approach to organising their delivery. Briefly, he features of appear to be as follows:

- Testing when ready – teachers are to decide when a level has been attained
- Shorter more focused and more appropriate tests
- Single level tests
- Externally set and marked, delivered twice a year
- "One way Ratchet" - never going back, only forward
- Age independent tests
- Used for accountability, retaining current performance standards.

Each of these elements will be discussed in turn.

Testing When Ready

In general, the notions of testing when ready and the close tie to teaching and learning are laudable. They fit within the context of Personalised Learning and Assessment for Learning. As such "Single Level Tests" could provide a useful stimulus to teaching

and learning. However, as described in ***Making Good Progress***, it is doubtful that they can fulfil that function. As a single level test, awarding a level, the test would generally show what a student could do but it would not be able at the same time to provide diagnostic information about the next steps since these would not be included in the test. Similarly, because it would have to cover the curriculum broadly at that level, and levels represent two years of teaching (on average), it could not identify the small next steps needed for personalised learning. It is perhaps significant that in the more recent announcements on Single Level Tests, the diagnostic function has not been mentioned. It may well have been quietly dropped.

The proposed assessment system is to be built on the foundation of the teachers' assessment of their students. Hence teacher assessment itself will need to be reasonably reliable and as valid as possible. The proposed method of ensuring this is through the provision of materials known as Assessing pupils' progress (APP). This is designed to improve learning outcomes by providing teachers with more effective assessment approaches. These include criteria and other materials. There is no moderation system envisaged at present – in effect this may be provided by the Single Level Tests, raising again the spectre of teaching to the tests, but now more narrowly defined tests. It would seem essential that the success of the APP process is evaluated.

Although the evaluation of the first run of the tests in December 2007 is not yet available, it is known that the results were unexpected. Sue Hackman, chief adviser on school standards at the Department for Children, Schools and Families (DCSF) wrote to schools and apologised for a delay in returning the results. This was said to be due to the marking and level-setting process revealing some unexpected patterns in the results. It seems to have been the case that the proportions of students achieving the levels for which they were entered were lower than expected. This may have been due to the nature of the tests (which were untried), to the level-setting process or to a mismatch between teachers' judgements of the attainment of a level and the standards required by the tests. All of these possibilities will require investigation and improvement in subsequent rounds of the pilot.

There is a further issue which relates to the concept of testing when ready. This can be a useful process, particularly if it is used formatively and incorporated into teaching and learning. However, its utility within a summative system may not be as apparent. The argument advanced in ***Making Good Progress*** is that success at one level will stimulate progress toward the next level, acting motivationally. This will need to be evaluated in practice. It may be that the levels are so far apart (they are intended to cover two years of development) that achieving one level may actually slow progress to the next, since it may be too distant a target. This is particularly a concern because of the 'one way ratchet' proposal. The achievement of a level and the knowledge that it cannot be removed may act to demotivate rather than motivate.

Shorter More Focused Tests

Making Good Progress states that the tests will be 'shorter and more focused.' Since there is a strong relationship between reliability and test length, there is unfortunate implication that the tests will have lower levels of reliability and reduced curriculum coverage. Paradoxically, ***Making Good Progress*** also makes clear that the proposed tests would be used for accountability purposes, with the levels awarded being retained for ever and reported. This means that the tests will need to have the characteristics of tests for accountability: high levels of reliability and validity.

In this context, the important aspect of reliability is the consistency of the decisions made. It there were two progress tests at the same level, what would be the percentage of students classified the same way on both occasions? For the tests to be shown to be useful, this needs to be considerably above chance levels.. This would need careful examination during development, as reducing the length of test inevitably leads to lower levels of reliability.

There is a distinct problem with conceptualising the reliability of the Single Level Tests. As with all notions of reliability, at its heart must be consistency and repeatability. The Single Level Tests system has several components which will all need to have good levels of reliability. It will require:

- The teacher judgements used to enter the students to themselves be reliable
- The entry decisions to be consistent between teachers
- The marking of the tests to be reliable
- The tests to give consistent decisions about students.

At present, none of these have been demonstrated.

In relation to the tests themselves, there will have to be demonstrations of decision consistency. Ideally, there would be parallel tests taken by the same individuals or exercises in which students took the same test twice. However, in a high stakes system with testing on demand, this will prove difficult.

In such circumstances, it is usual to fall back on internal consistency measures such as Cronbach's alpha. These have the advantage of usually being calculable from a single test administration, but as such they can give high estimates of reliability since they do not allow for differing test circumstances that arise from testing on two occasions. There is a further issue for the single level tests in that internal consistency measures are most useful when there is no restriction in the range of candidates' abilities. By there very nature, Single Level Tests will, if the system is working, only be taken by a narrow band of candidates, who have just achieved the level. In such circumstances, procedures for establishing the reliability of the Single Level tests remain to be developed.

A second aspect of the 'shorter more focused' is curriculum coverage. In the current National Curriculum tests considerable efforts are made to include as wide a representation of the curriculum as is practically possible in a written test. This is essential for demonstrations of validity. Moreover, the annually changing tests mean that, over time the tests have even wider coverage. In writing for example, different text types/genres are sought from children each year and, within the test each year, two different tasks are required. Hence, reducing the length of the tests could also reduce the validity of the test.

Single Level Tests

The meaning of the phrase 'Single Level Tests' needs some exploration. Behind the proposals of Making Good Progress, there seems to be a naïve view that questions can be written at a single level, derived from the level descriptors and these will have comparable difficulty. It is not the case that the levels of the National Curriculum are, in practice, as even and well ordered as the underlying model would suggest Taken to its extreme, it is sometimes thought that a single level test could be constructed by having material drawn from the level descriptor at that level. Candidates would then be expected to answer a set proportion of this correctly. There have been examples of such systems which have been constructed with these principles and in which the consequence has been very low pass rates.

Happily, it does seem that this early concept naivety has not been continued and there are now measures in place for setting and maintaining the standard of the tests from one administration to the next session. There does though (as with so much) need to be continuing research and development of this aspect.

Externally set and marked, delivered twice a year

The delivery proposals for Single Level Tests are immediately a compromise with the underlying conception. The notion of "testing when ready" should be just that, allowing testing to take place at any time. The modern solution would be computer delivered tests, possibly tailored and certainly individualised. This could be drawn from a central bank. However there are problems delivering such a system at present. The first is the high-stakes nature of the tests. This means that the delivery system must be 100 per cent reliable. This is currently achievable (more or less) in systems using dedicated test centres, but has not been demonstrated to be achievable in the schools IT environment. This is particularly the case for primary schools, which have less advanced systems and poor technical support. A second issue is the nature of the tests. The tests contain questions requiring both short and long answers, some of which may not be amenable to computerised marking but which require human judgement. This is currently easier to source and organise as a series of specific events rather than as an ongoing on-demand operation.

For these reasons, the initial proposal is for testing exercises at two times each year. However, even this may not be sustainable in the long term. The high-stakes nature of the tests means that the tests or the items cannot be re-used. This leads to a high

cost of development for a new set of tests twice each year, some of which (at the higher levels) may be taken by very few students. Currently, the tests are being developed using sound, classical processes to ensure good levels of validity and the maintenance of standards as well as possible. However, the costs of this may lead to a search for cheaper alternatives, and the challenge will be to ensure that the development process remains sufficiently rigorous to provide dependable results.

One-way Ratchet

There are further concerns about the 'one-way ratchet'. Its underlying assumption seems to be that children's learning is an ordered progression and that movement is always forward. This is not in fact the case, and children can decline in terms of skills or knowledge. It is therefore useful to have later checks that a level previously achieved has been maintained. If this is not the case, we do not believe the "one way ratchet" should be implemented.

This issue may interact with that of the reliability of the test. If the decision consistency of the tests at a given level is low, then a large proportion of candidates could be misclassified as achieving the level when they should not. If this is coupled with the 'one way ratchet', the misclassification would become enshrined, possibly being harmful to such children's progress as they would be being treated (and taught) as if they were at a higher level than was actually the case.

It would seem sensible that that the 'one way ratchet' is abandoned and that the system allows for re-testing of doubtful cases so that high levels of certainty are achieved and so that misclassification is minimised. A useful refinement would be to have a system in which there are three levels of outcome: level X awarded; level X not awarded; and a band of uncertainty in which a retest is advised in the following test round. Hence teachers could report only success which is assured to a high probability, requiring pupils with scores in a defined range of uncertainty to be retested. However there is no indication that this will become part of the system

Age Independent Tests

An interesting part of the proposals is that the same tests will be used whatever the age of the student, provided that they are at that level. Although this concept is used in some graded test systems, such as music examinations, it is unusual in educational assessment. This means, for example that the reading test at level 4, must be able to be taken by a very able 8-year-old and, equally well, by a 14-year-old who is struggling. The content and format must be equally accessible and attractive to both students. This puts a considerable load on the test developers to produce such material, if indeed it is possible. For mathematics there are further problems. Although ostensibly related to the same levels, the Key Stage 2 and Key stage 3 programmes of study are in fact different. For Single Level Tests at a level which spans both key stages, decisions have to be taken as to what content is to be included. Should algebra for example be included in mathematics tests at level 5, when it is not part of the Key Stage 2 curriculum, but primary children may take the test? In effect,

decisions like these may lead to a taught curriculum based on the tests rather than the official National Curriculum. If one of the purposes of Single Level Tests was to reduce narrowing of the curriculum, it may in fact have the opposite effect.

It remains to be established whether it is possible to have tests which are suitable for a very wide age range, providing sufficient motivation and support to younger children yet not being condescending to older students. The curriculum effects also remain to be established.

Used for Accountability, Retaining Existing Performance Standards

It is the classic dilemma of assessment development that policy makers require change and new systems while at the same time also requiring that the standards awarded must be equivalent to those of the existing system. This is assessment nonsense but a political reality. This dilemma occurred early in the Single Level Tests pilot. The initial specification for the first, December 2007, tests stated that the standard set for the tests should be based on a "secure" operation at a given level, not those of the existing National Curriculum tests which are set to the threshold of the level. The teacher assessment to identify those ready for the tests would similarly be based on a secure achievement of the level. This was in one sense eminently reasonable, since it fitted with other aspirations of the system. However, another part of the proposals conflicted with this. Even during the pilot stages, the award of levels from Single Level Tests to students was to override their National Curriculum Test results. Hence these are the results to be reported to parents and to be included in performance tables. Using untried tests being piloted to supplant results from an existing established system seems curious, but the Government remains adamant that this should be the case. However, if the cut-scores used were "secure", that is well into the range of a level, this is effectively a different standard from the threshold criterion of the existing data. While in the pilot, this might not make much difference, in a fully rolled-out system, the effect could be to lower the proportion of students achieving levels since the standard required had effectively been raised.

This realisation dawned a little late, after the first round of the pilot tests in December 2007 and in the following February, it was announced that the standards used for future tests would be those of the existing national curriculum tests, attempting to give continuity to the meaning of levels and, perhaps more importantly, to performance tables. There are now plans for equating the existing new Single Level Tests to the existing National Curriculum Tests for the current and future rounds of the pilot.

Conclusion

The pilot of Single Level Tests as with all of ***Making Good Progress*** remains at an early stage. It is too early to say whether it will be successful. As with many innovations, success would have several facets: it would achieve the political objectives; it would have the trust of teachers and the public; and it would achieve the educational measurement criteria of reliability and validity. These are all stringent requirements and somewhat in conflict with each other. However, at present, it does

remain a pilot and provided evaluation evidence is generated and the lessons are learned, it should be possible for a Single Level Test system to evolve. However, it is unlikely to be, and should not be expected to be, a complete realisation of the initial concept.

References

Department for Education and Skills. 2007. *Making good progress: How can we help every pupil to make good progress at school?* London: DfES.

Harlen, Wynne. 2007. *The quality of learning: assessment alternatives for primary education. Primary review research survey 3/4.* Cambridge: Esmée Fairbairn Foundation.

Newton, P.E. 2007. Clarifying the purposes of educational assessment. *Assessment in education: Principles, Policy & Practice* 14, no. 2: 149–170.

Statistics Commission. 2005. *Measuring standards in English primary schools. Report no.23*. London: Statistics Commission.

Tymms, P. 2004. Are standards rising in English primary schools? *British Educational Research Journal* 30, no. 4: 477–494.