



Qualifications and
Curriculum Authority

Single level tests

*Report of the first three test sessions: December 2007, June 2008 and
December 2008*

July 2009

QCA/09/4303

Contents

Executive summary.....	3
National curriculum tests and single level tests.....	5
National curriculum tests	5
Single level tests.....	6
The test model	7
Test development.....	9
Timelines	9
Requirements for single level tests and generic issues around test development...	11
Test specifications	12
Single level tests in reading	12
Single level tests in writing.....	15
Single level tests in mathematics.....	18
Item evaluation and the first pre-test	22
Standard setting: Level setting for single level tests for December 2007	24
Level setting for single level tests for June 2008.....	29
Initial findings from test equating for the June 2008 single level tests	40
Level setting for the June 2008 single level tests	48
Results from the June 2008 single level tests session	54
Interpreting the findings from the June 2008 test session	55
The December 2008 test session	57
Final level setting.....	58
Taking forward the development of single level tests.....	60
Appendix A – Equating design for the June and December single level test pre-test	62
References.....	65

Executive summary

QCA was commissioned to develop and pilot single level tests as part of the Making Good Progress initiative. This report provides a factual account of the first three single level test sessions (December 2007, June 2008 and December 2008), setting out the test model, the test development process, the standard setting processes and the outcomes.

The single level test approach was innovative, with no direct comparators in other systems of national assessment. Single level tests also differed from national curriculum tests in a number of key aspects. For national curriculum tests, there were two suites of tests, with the construct assessed by the key stage 2 tests being the key stage 2 programme of study and the construct assessed by the key stage 3 tests being the key stage 3 programme of study. The single level tests model required a single suite of tests, covering all levels and spanning across both programmes of study. There was, for single level tests, no common programme of study which could serve as the basis for test development.

The purpose of the pilot was to generate evidence on three linked issues, relating to the nature of the tests:

- (a) are there common knowledge and common skills (a common test construct) which underlie both the key stage 2 and key stage 3 programmes of study and is it possible to develop a single suite of tests which assesses that common test construct?
- (b) if such tests can be developed, how would standards be set on them?
- (c) Is it possible to develop age-independent tests, which can be sat by pupils at any point from year 3 through to year 9?

The first test session, held in December 2007, was groundbreaking. Highly experienced test developers hypothesised a common test construct for each subject across the two key stages, consisting of the knowledge and skills they judged pupils working at particular levels would exhibit, independent of age. They then wrote test items based on that hypothesised construct. Teachers were asked to enter pupils for single level tests in December 2007 when they judged that the pupils were working securely at the level of the test and the cutscore for the test was set at "secure" performance, to mirror that teacher judgement. The outcomes from the live tests were unexpected, with tests at the lower levels functioning quite well, but there were issues at the higher levels and between pupil performance at the two key stages.

Single level tests for the June 2008 test session and the December 2008 test session were developed concurrently and designed so that the cutscore could be set at

"threshold" performance, rather than "secure" performance, to align them with the standards of national curriculum tests. On the basis of the findings from December 2007, a large-scale pre-test was designed, to provide a rigorous trial of the tests before they were sat by pupils in pilot schools. Test developers reviewed script evidence from December 2007 and, in the light of that, re-hypothesised and wrote items based on a revised common test construct. Approximately 35,000 pupils in schools outside the Making Good Progress pilot sat these tests, as a pre-test. Psychometricians then carried out a detailed analysis of the pre-test data and this began to explain the findings from the December 2007 test session.

The main findings were that pupils from key stage 2 outperformed those from key stage 3 and this was mainly because the single level test model, with separate tests and targeted items for each level, fitted for key stage 2, but not for key stage 3. Single level tests did not seem to fit with the curriculum model for key stage 3, where it did not seem to be possible to target items at particular levels. Analysis by senior markers confirmed that there were differences in the nature of pupil performance at key stage 2 and key stage 3 at the common levels.

In the light of the findings from the June 2008 test session, key stage 3 pupils were withdrawn from the single level tests pilot, although they continued to participate in the Making Good Progress initiative. The December 2008 single level tests were sat by pupils from key stage 2 only and the outcomes were encouraging.

On the basis of the evidence gathered, four clarifications of the test model were agreed:

- (i) the construct being assessed was the key stage 2 programme of study;
- (ii) the standard for the tests should be the current key stage 2 standard;
- (iii) age independence was to be conceptualised in terms of test entry decisions, so a pupil could be entered for a single level test at any point during years 3-6;
- (iv) tests confirmed a teacher assessment judgment by providing an independent measure of a pupil's attainment with respect to the national curriculum.

On the basis of the research evidence generated during the pilot and in the light of the clarifications to the test model, QCA is well-placed to take forward the development and further piloting of single level tests.

National curriculum tests and single level tests

National curriculum tests

National curriculum tests assess a pupil's level of attainment with respect to the national curriculum. Up to the May/June 2008 national curriculum test session, there were two sets of national curriculum tests, one for key stage 2 pupils and one for key stage 3 pupils. In October 2008 the statutory tests for key stage 3 pupils were discontinued. Pupils in maintained schools in England took national curriculum tests at the end of key stage 2 (usually year 6, age 10/11) to assess the key stage 2 programme of study, and at the end of key stage 3 (usually year 9, age 13/14, to assess the key stage 3 programme of study).

End-of-key-stage national curriculum tests are available in three subjects: English, mathematics and science. The English test comprises a reading component and a writing component and a separate level is reported for each. The marks for each component are aggregated to give an overall result for English. The mathematics test comprises a calculator paper, a non-calculator paper and a mental mathematics test. Again, the marks from these papers are aggregated to give an overall level for mathematics. The science test consists of two papers, for which the marks are combined to give a level. End-of-key-stage test papers cover a range of levels:

Key stage 2

English	Levels 3–5
Mathematics	Levels 3–5
Science	Levels 3–5

Key stage 3

English	Levels 4–7
Mathematics	Levels 3–5, 4–6, 5–7, 6–8
Science	Levels 3–6, 5–7

Pupils sit end-of-key-stage tests once, at the point when they have completed the relevant programme of study. This is normally in year 6 and year 9. Pupils cannot re-sit the tests.

Single level tests

Single level tests were available in three subjects for all three test sessions: English reading, English writing and mathematics. At the beginning of the pilot it was anticipated that reading and writing would be aggregated into a total score for English. For all three test sessions, a single level test was aimed at a single national curriculum level and reported in terms of whether or not a pupil had demonstrated attainment at that level.

Single level tests were available twice a year (with test sessions in June and December) and a pupil could retake a single level test at a future test round, if judged appropriate by their teacher.

For the December 2007 test session, the following 12 tests were available:

English reading: Level 3, 4, 5, 6

English writing: Level 3, 4, 5, 6

Mathematics: Level 3, 4, 5, 6

For the June 2008 test session, the following 18 tests were available:

English reading: Level 3, 4, 5, 6, 7, 8

English writing: Level 3, 4, 5, 6, 7, 8

Mathematics: Level 3, 4, 5, 6, 7, 8

For the December 2008 test session, the following 12 tests were available:

English reading: Level 3, 4, 5, 6

English writing: Level 3, 4, 5, 6

Mathematics: Level 3, 4, 5, 6

For the December 2007 and June 2008 single level test sessions, pupils from years 3 to 9 were eligible for test entry. For the December 2008 single level test session, entries were restricted to pupils in years 3 to 6. Pupils from key stage 3 were removed from the single level test strand of the Making Good Progress pilot following the discontinuation of statutory national curriculum testing for key stage 3.

Pupils who did not achieve the level in a single level test were eligible to re-sit.

The test model

The remit, as set out in the project initiation document for single level tests, was to 'develop new style tests in reading, writing and mathematics targeted at single national curriculum levels with level outcomes in mathematics, English reading and English writing.' The criteria for test development were that:

- each test would contain only questions targeted at the level being tested
- each test would be of 50 minutes' duration
- the tests would look as similar as possible to current national curriculum tests. The reading tests would therefore consist of linked texts with questions (but no Shakespeare paper), the writing tests would comprise a shorter and a longer task, and the mathematics tests would comprise a calculator and a non-calculator paper (but no separate mental mathematics test, as these skills would be tested through the papers).

The following tests were developed for the December 2007 test session:

- tests in reading, covering levels 3, 4, 5 and 6 separately
- tests in writing, covering levels 3, 4, 5 and 6 separately
- tests in mathematics, covering levels 3, 4, 5 and 6 separately.

There were two major changes to the test model for June 2008. First, the tests themselves were restructured. For December 2007, all items in a test were judged to be appropriate for a pupil working securely at the level of that test. However, it had already been decided that single level tests should carry forward the standards of national curriculum tests, so that, for instance, a level 3 in a single level test would be 'worth' the same as a level 3 in a national curriculum test. In a national curriculum test, a pupil who demonstrates performance at the threshold of a level is awarded that level, so single level tests were adapted to this model.

There were two possible ways of implementing this at a practical level. The model in which all items in a single level test were targeted at the level of the test could have been maintained, but with a cutscore set at a very low point in the mark distribution.

Alternatively, a test could have been developed in which a proportion of the items were targeted at the level below that of the test, with the remainder of the items at the level of the test. The cutscore could then be set at a higher point in the mark distribution.

The first option was not considered to be defensible, because a very low cutscore would have meant that a pupil could, theoretically, have gained marks on the tests in a wide variety of ways and apparently at random. The latter option was potentially the technically more robust alternative and, for that reason, tests were developed in which approximately 40 per cent of the marks would be available for attainment at the level below that of the test, with the remaining 60 per cent of the marks for attainment at the level of the test. In the case of mathematics, this resulted in tests for consecutive levels having some items in common. For reading, the requirements of the test model were met by increasing the demands/difficulty of the questions within a test paper. The intention was that cutscores could be set at between 40 and 50 per cent of the marks.

This meant that tests no longer contained only questions targeted at secure knowledge of the level being tested, but questions targeted across the threshold of the level being tested. This was consistent with the intention underlying single level tests: to provide a pupil with an opportunity to demonstrate performance at a particular national curriculum level.

The second change was a response to evidence, from the December 2007 tests, of relatively high non-completion rates, particularly for the higher level tests. The time allocation for the higher level tests was increased and details are given in Table 1 below.

Table 1: Revised time allocations for single level tests from June 2008

Subject	Level(s)	Time allocation
Reading	3–5	40 minutes plus 10 minutes reading time
	6	50 minutes plus 10 minutes reading time
	7–8	60 minutes plus 10 minutes reading time
Writing	3–5	50 minutes
	6	60 minutes
	7–8	70 minutes
Mathematics	3–5	25 minutes for each of paper 1 and paper 2
	6	30 minutes for each of paper 1 and paper
	7–8	35 minutes for each of paper 1 and paper 2

The December 2008 tests were developed concurrently with the June 2008 tests and, structurally, mirrored the June 2008 tests.

Following the June 2008 single level test session, further changes were made to the test model. On 14 October 2008, the Secretary of State announced the discontinuation of statutory national curriculum testing at key stage 3 and halted the key stage 3 strand of single level tests. Pupils from key stage 3 were therefore no longer eligible for test entry. Pupils in key stage 2 (years 3–6) continued to be eligible for test entry and tests were available at levels 3–6.

Test development

Development of the 12 separate single level tests for the December 2007 test session had a truncated timetable of 8 months. By way of comparison, the timescale for the development of the key stage 2 national curriculum tests is approximately 26 months.

Because of the timescales, it was not possible to procure test development agencies to develop the December 2007 tests, so QCA led on the test development process, supported by highly experienced consultants and test development agencies. For English reading and writing, the team for test development for the December 2007 single level tests comprised: two senior test developers with recent test development agency experience; one former senior marking programme leader with curriculum and assessment expertise and one English curriculum and assessment expert.

For the mathematics tests, the team for test development for the December 2007 single level tests comprised: two senior test developers, one with very recent test development agency experience and two mathematics curriculum and assessment experts.

The QCA procured test development agencies to develop the June 2008 and December 2008 single level tests.

Timelines

Table 2 below sets out the timeline for the development of the December 2007 single level tests.

Table 2: Timescale for the development of single level tests: December 2007 test session

Timescale	Activity
May/June 2007	Recruitment of test development consultants and identification of pre-tested, but unused, national curriculum test items.
July 2007	Review and, if appropriate, adaptation of test items and mark schemes. Production of new test items, if appropriate. Test review group meetings. Resolution meeting (mathematics).
August 2007	Design of test papers. Further review of and amendments to test items.
September 2007	Recruitment of schools for, and carrying out of, informal trial of items. Test review groups. Resolution (English reading and writing). Review of scripts from the informal trial and, if appropriate, amendments to the test papers/mark schemes in the light of pupil performance.
October 2007	Final revisions to the test papers/mark schemes. Project Board sign-off of test materials. Proofing.
November 2007	Printing of the question papers.

Single level tests for the June and December 2008 test sessions were developed concurrently. The process started at the beginning of November 2007 and was completed at the end of April 2008, with details given in Table 3 below.

Table 3: Timeline for the development and production of the June and December 2008 single level tests

Activity	Start	Finish
Development of test specification	07/11/07	28/01/08
Identification of item writers/pre-test agency/design agency	7/11/07	21/12/07
Item writing and mark scheme development	26/11/07	03/01/08
Informal trialling	10/12/07	08/02/08
Pre-test 1 (from definition of specification to resolution meeting)	03/12/07	01/04/08
Pre-test 2		31/07/08
Handover 1 of test papers	07/04/08	
Handover 2 of test papers	21/04/08	
Handover to print	28/04/08	

Requirements for single level tests and generic issues around test development

As noted above, the specification for test development made the following high-level requirements for single level tests, which mirrored the requirements for end of key stage testing.

'Single level tests ... must:

- generate results that provide a valid measure of the required knowledge, skills and understanding as defined by the national curriculum orders
- generate results that provide a reliable measure of pupil performance
- generate results which provide comparability of standards
- minimise bias, differentiating only on the basis of all pupils' ability to meet national curriculum requirements
- deliver a manageable system of assessment.'

Materials developed as part of the assessing pupils' progress (APP) initiative were used to support test development for single level tests. This was the case across all subjects, but was most significant for the writing tests, where the generic mark scheme used for national curriculum testing was adapted for single level testing, using the APP criteria as its basis. The APP criteria were used by teachers in schools in the Making Good Progress pilot, to support their teacher assessment judgments.

Test specifications

For national curriculum tests, the construct assessed by key stage 2 tests is the key stage 2 programme of study and the construct assessed by key stage 3 tests was the key stage 3 programme of study. These programmes of study form the basis for test specifications, which set out the precise requirements for key stage 2 and key stage 3 tests.

The requirement for a common single level test at levels 3, 4 and 5 across the two key stages meant that the construct being assessed by single level tests could not be the key stage 2 or the key stage 3 programme of study. Test developers, therefore, needed to use their professional judgment to develop tests which they felt could be appropriate for piloting with pupils in both key stages.

For the June 2008 and December 2008 test sessions, statements of the functional requirements for tests in each subject were developed, as set out below. It should be noted that these statements of the functional requirements for tests were produced when single level tests were being developed for pupils in both key stage 2 and key stage 3, so refer to years 3-9 and levels 3-8.

Single level tests in reading

The functional requirements for single level tests in reading for June and December 2008 are given in Table 4 below.

Table 4: Functional requirements for single level tests in reading

Target year group	Years 3 to 9.
Levels assessed	All levels from 3 to 8.
Legal status	Optional for local authority–maintained schools taking part in the Making Good Progress pilot.
Delivery mode	Written tests printed and distributed to schools.
Pupil eligibility	Pupils whom the teacher assesses as working securely at the level of the test.
Model of assessment	Reading booklet containing two texts for levels 3 and 4 and at least three texts for level 5 and above. Separate booklet for each level. Reading answer booklet containing questions based on the reading booklet for each level.
Timing	For levels 3 to 6, 40 minutes plus 10 minutes of reading time, total time 50 minutes. For levels 7 and 8, 60 minutes plus 10 minutes of reading time, total 70 minutes.
Available marks	Each test should have up to 40 marks, dependent upon level: <ul style="list-style-type: none"> • at levels 3 and 4, no question will be allocated more than three marks and there will be no more than two three-mark questions per test • at levels 5 to 8, four-mark questions will occur; one of these will be used at level 5, a maximum of two at level 6 and at least two at levels 7 and 8.
Curriculum coverage	The questions included in the tests will be drawn from across the full range of the reading programme of study. Pupils should be prepared to answer questions on any texts from the ranges specified in the programmes of study.
Question types	The tests will be keyed to the following assessment focuses (AFs) based on the level descriptions: AF1: use a range of strategies, including accurate decoding of text to read for meaning

	<p>AF2: understand, describe, select or retrieve information, events or ideas from texts and use quotation and reference to text</p> <p>AF3: deduce, infer or interpret information, events or ideas from texts</p> <p>AF4: identify and comment on the structure and organisation of texts, including grammatical and presentational features at text level</p> <p>AF5: explain and comment on the writers' use of language, including grammatical and literary features at work and sentence level</p> <p>AF6: identify and comment on writers' purposes and viewpoints and the overall effect of the text on the reader</p> <p>AF7: relate texts to their social, cultural and historical contexts and literary traditions.</p> <p>The tests should be based on a small number of relevant stimulus texts that are accessible to the reader who is operating at that level in reading, but relevant to a wide range of ages (7 to 14) and levels of maturity.</p> <p>Texts should represent a selection from the range of reading material described in the national curriculum English programmes of study, including those from other cultures and traditions, and over a number of successive tests reflect coverage of the full range of materials described in the programmes of study. Texts may be literary or related to other subjects in the national curriculum or to the world beyond.</p> <p>The texts should be linked by a common theme, offer a cohesive focus across the texts and provide an opportunity for overview questions.</p> <p>Texts should be of good quality and unlikely to have been experienced by a significant number of pupils.</p> <p>There should be a range of question types, including those requiring an extended written response (see section on 'available marks' above).</p> <p>Each of the questions should relate to one specific reading assessment focus. Questions should be varied and appropriate to the topics and AFs. The balance of marks for each AF will vary from year</p>
--	--

	<p>to year, but the test should include questions on reading AFs 2–7. At levels 3 and 4, greater weighting should be put on AFs 2 and 3 and at levels 5 and above on AFs 4, 5, 6 and 7.</p> <p>Questions on test papers should provide accessible starting points.</p> <p>Approximately 40 per cent of the marks will be from questions below the level being assessed.</p>
Outcomes/ deliverables	<p>Reading tests at each of levels 3, 4, 5, 6, 7 and 8 and associated mark schemes. Each level test will consist of a 12-page reading stimulus booklet and a booklet of questions in which pupils write their answers.</p> <p>The mark schemes booklet will be used by the external markers. The mark schemes must provide sufficient and clear guidance for markers to mark with consistency and accuracy, and for the marking of the tests to be effectively standardised. The reading test development agency will be required to work with the marking agency to ensure that marker training materials accurately reflect the intention of the mark scheme.</p>

Single level tests in writing

The functional requirements for single level tests in writing for June and December 2008 are given in Table 5 below.

Table 5: Functional requirements for single level tests in writing

Target year group	Years 3 to 9.
Levels assessed	All levels from 3 to 8.
Legal status	Optional for local authority–maintained schools taking part in the Making Good Progress pilot.
Delivery mode	Written tests printed and distributed to schools.
Pupil eligibility	Pupils whom the teacher assesses as working securely at the level of the test.

Model of assessment	<p>Two writing tasks at each level:</p> <p>shorter (assessing ability to write concisely/precisely)</p> <p>longer (test of extended writing).</p> <p>Spelling will be assessed over both tasks.</p>
Timing	<p>Writing tests comprise two tasks.</p> <p>For levels 3, 4, 5 and 6, 50 minutes will be allowed.</p> <p>For levels 7 and 8, 70 minutes will be allowed.</p>
Available marks	<p>The tasks will be made level-appropriate through the prompts and the specific requirements of the writing and will be defined by customisation of the generic mark scheme for single level tests.</p> <p>The total marks available will be 27 for levels 3 and 4 (where handwriting is assessed and spelling carries a greater weight) and 23 for level 5 and above.</p>
Curriculum coverage	<p>The questions included in the tests will be drawn from across the full range of the writing programme of study. Pupils should be prepared to write in any form covered within the programme of study.</p>
Question types	<p>The tests will be keyed to the following assessment focuses based on the level descriptions:</p> <p>AF1: write imaginative, interesting and thoughtful texts</p> <p>AF2: produce texts which are appropriate to task, reader and purpose</p> <p>AF3: organise and present whole texts effectively, sequencing and structuring information, ideas and events</p> <p>AF4: construct paragraphs and use cohesion within and between paragraphs</p> <p>AF5: vary sentences for clarity, purpose and effect</p> <p>AF6: write with technical accuracy of syntax and punctuation in phrases, clauses and sentences</p> <p>AF7: select appropriate and effective vocabulary</p> <p>AF8: use correct spelling.</p> <p>There will be one longer and one shorter task. There should be</p>

	<p>sufficient differentiation between the longer and shorter writing tasks in terms of purpose form, context and level of formality. These tasks should provide opportunities for pupils to respond creatively and imaginatively.</p> <p>Both tasks should cover, over time, the range of purposes and forms detailed in the national curriculum programmes of study.</p> <p>There will be a planning sheet for the longer writing task but planning will not be marked.</p> <p>Spelling will be assessed across the longer and shorter writing task. Handwriting will be assessed at levels 3 and 4 only.</p> <p>Tasks should be clearly and concisely worded, offering an appropriate level of contextual support while allowing opportunities for pupils to interpret the tasks. They should be presented clearly and attractively.</p> <p>Mark schemes should relate to the writing AFs and draw them into three strands for the longer writing task:</p> <p>composition and effect (AF1–2)</p> <p>text structure and organisation (AF3–4)</p> <p>sentence structure and punctuation (AF5–6)</p> <p>and three strands for the shorter writing task:</p> <p>composition and effect (AF1–2)</p> <p>sentence structure, punctuation and text organisation (AF4–6)</p> <p>spelling (AF8)</p>
<p>Outcomes/ deliverables</p>	<p>Final products: Writing tests at each of levels 3, 4, 5, 6, 7 and 8 and associated mark schemes. Each writing test will consist of two writing task prompts and a booklet in which the pupils write their answers.</p> <p>The mark scheme will be used by the external markers. The mark schemes must provide sufficient and clear guidance for markers to mark with consistency and accuracy and for the marking of the tests to be effectively standardised.</p>

Single level tests in mathematics

The functional requirements for single level tests in mathematics for June and December 2008 are given in Table 6 below.

Table 6: Functional requirements for single level tests in mathematics

Target year group	Years 3 to 9.
Levels assessed	All levels from 3 to 8.
Legal status	Optional for local authority–maintained schools taking part in the Making Good Progress pilot.
Delivery mode	Written tests printed and distributed to schools.
Pupil eligibility	Any pupil from year groups 3–9 can be entered for a single level test once the teacher has established, through the use of the APP assessment criteria, that the pupil is working securely at the level of the test they are entered for and has progressed and is working at a level that is one or more levels higher than their most recent single level test or national curriculum test assessment.
Model of assessment	Two written papers: paper 1 (non-calculator) paper 2 (calculator). There is no separate mental mathematics element to single level tests.
Timing	Levels 3–6: pupils are allowed 25 minutes to complete each of the written papers. Both papers are to be taken in a single 50-minute session. Levels 7 and 8: pupils are allowed 35 minutes to complete each of the written papers. Both papers are to be taken in a single 70-minute session. There will be no mental mathematics element for the June and December 2008 single level tests.
Available marks	The number of marks available on the calculator and non-calculator paper should be the same within a level (or within one or two marks of each other).

	<p>Each paper should total 20–25 marks (for a 50-minute test).</p> <p>The total number of marks available may vary by level dependent upon the demand of the questions and time duration of the test.</p> <p>For example, at levels 7 and 8 the marks available should be adjusted to account for the longer time permitted than for levels 3–6.</p>
<p>Permitted equipment</p>	<p>Papers 1 and 2:</p> <p>a ruler (showing centimetres and millimetres)</p> <p>an angle measurer or protractor</p> <p>a pair of compasses</p> <p>tracing paper</p> <p>a mirror.</p> <p>Paper 2 only:</p> <ul style="list-style-type: none"> • a calculator. <p>For levels 3–6, there will be no specified type of calculator required by pupils. They will be instructed to use the calculator they use as part of usual classroom practice.</p> <p>At levels 7 and 8, the requirement to use a scientific or graphic calculator will be defined.</p>
<p>Formulae permitted</p>	<p>Any formulae that pupils are not required to remember must be provided in a standard format at the start of the papers.</p>
<p>Balance of marks</p>	<p>The number of marks available for each national curriculum attainment target for each test should be in the following ratios.</p> <p>For levels 5 and above:</p> <p>Ma2 (number and algebra), Ma3 (shape, space and measures) and Ma4 (handling data) should be in the ratio 9:4:3.</p> <p>For levels 3 and 4:</p> <p>Ma2 (number and algebra), Ma3 (shape, space and measures) and</p>

	<p>Ma4 (handling data) should be in the ratio 5:2:1.</p> <p>Across both papers at all levels, using and applying mathematics (UAM) (Ma1) marks should feature in approximately 20 per cent of the marks available (e.g. there should be UAM features in 3-4 of every 20 marks).</p> <p>These marks should be spread equally across the three strands of UAM and, as far as possible, the proportions should reflect the ratios of the attainment targets.</p> <p>As far as possible, UAM items should be designed as such, rather than writing items and retrospectively identifying those that seem to most assess UAM.</p> <p>The marks at level should be awarded for performance consistent with the demand indicated by the level descriptions in the APP guidelines and the national curriculum. <i>Note: these guidelines are minimal for levels 6, 7 and 8 and very often reflect key stage 2 practice rather than key stage 3.</i></p> <p>Questions on test papers should provide accessible starting points and progression through the paper. Questions should progressively increase in demand through the paper, although it is acknowledged that question order may need to take account of other practical constraints, such as the layout of questions on the paper.</p>
<p>Curriculum coverage</p>	<p>The questions included in the tests will be drawn from across the full range of the mathematics programmes of study.</p> <p>Ma2: Number and algebra:</p> <ul style="list-style-type: none"> • using and applying number and algebra • numbers and the number system • calculations • solving numerical problems • equations, formulae and identities

	<ul style="list-style-type: none"> • sequences, functions and graphs. <p>Ma3: Shape, space and measures:</p> <ul style="list-style-type: none"> • using and applying shape, space and measures • geometrical reasoning • transformations and coordinates • measures and construction. <p>Ma4: Handling data:</p> <ul style="list-style-type: none"> • using and applying data • specifying the problem and planning • collecting data • processing and representing data • interpreting and discussing results.
<p>Question types</p>	<p>A range, including:</p> <ul style="list-style-type: none"> • calculation • questions requiring the application of mathematical processes in contexts of Ma2, Ma3 and Ma4 • questions linking sections of the programme of study • questions requiring pupils to explain/justify their answers using mathematical reasoning • questions requiring pupils to produce unsupported solutions to multi-step problems • questions drawn from both real-life and mathematical contexts • questions requiring pupils to select the appropriate information

	<p>needed to solve a problem</p> <ul style="list-style-type: none"> • questions requiring pupils to determine the appropriate units and accuracy for their answer interpreting calculator outputs appropriately.
Question contexts	<p>Questions should make use of real-life and mathematical contexts. The contexts of questions should engage pupils whatever their age, cultural or social background or life experiences, and reflect common life experiences of 7- to 14-year-old pupils in schools in England.</p>
Outcomes/ deliverables	<p>Final products: single level mathematics tests at each of levels 3, 4, 5, 6, 7 and 8 and associated mark schemes. Each test will consist of two booklets in which the pupils write their answers.</p> <p>The mark scheme booklet will be used by external markers, so the mark schemes must provide sufficient and clear guidance for markers to mark with consistency and accuracy, and for the marking of the tests to be effectively standardised.</p>

Item evaluation and the first pre-test

National curriculum tests

For national curriculum tests, the process of item writing generally begins with initial versions of questions being drafted. These draft items are taken into schools for informal trialling and, on the basis of this, are amended and further developed. Once the items have reached an appropriate stage in their development, they are put together into coherent tests. These tests are checked against the test specifications to ensure that they are valid in terms of the content and skills being assessed. These tests are then taken to the first pre-test.

The focus of the first pre-test (pre-test 1) is on the performance of items. The tests are marked and then subjected to both qualitative and quantitative review, with the quantitative review generating evidence about the difficulty of the items (facilities) and the extent to which the items discriminated between candidates of different abilities (discriminations). On the basis of pre-test 1, items are reviewed and can be taken out of the test if they do not perform as desired.

Throughout the test development process, test materials are subjected to expert review. Test review groups (TRGs) are established groups of experts – teachers, local authority subject specialists and staff from higher education who are involved in teacher training – who review national curriculum test materials. TRG members comment on the extent to which the test materials:

- used appropriate language
- were interesting and motivating to pupils
- were valid interpretations of the programmes of study
- were pitched at an appropriate level of difficulty
- were illustrated appropriately
- were supported by a comprehensive mark scheme that recognises and rewards pupils' responses appropriately
- required advice on the suitability of materials from a cultural reviewer
- were accessible to pupils with a variety of special needs with the minimum of modification.

Once a final form of the test has been developed, it is taken to a second pre-test (pre-test 2). The purpose of pre-test 2 is to generate evidence for test equating purposes as part of the standard-setting process.

Single level tests

The process of item evaluation and pre-testing for single level tests necessarily differed from that for national curriculum tests, given the timescales for the development of the tests for that session.

For the December 2007 test session, given that there was not sufficient time to subject the tests to either a first or second pre-test, the tests were developed on the basis of qualitative professional judgment and expert review. The evidence for this was provided by an informal trial, involving approximately 100 pupils from primary and secondary schools sitting a test in each subject at each level. Test resolution meetings were then held and final amendments were made to the question papers and mark schemes.

For the June 2008 and December 2008 single level tests the test development process largely followed that used for national curriculum tests, although with shorter timescales.

TRGs were established for these rounds of single level tests, with the same remit as for national curriculum tests. A formal first pre-test was carried out by the test development agencies for the June 2008 and December 2008 single level tests, followed by a very large second pre-test (see "Level setting for single level tests for June 2008" below).

Standard setting: Level setting for single level tests for December 2007

For national curriculum tests, two approaches to standard setting are taken: test equating and script scrutiny. The purpose of these approaches is to recommend a series of cutscores on a test that are equivalent to the series of cutscores on a previous test.

For national curriculum tests, the first method of standard setting is test equating. This is a statistical approach that allows adjustment of the cutscores in a test to take account of its difficulty: the cutscores for a test will be lower than the previous test if the test was more difficult and higher if it was easier. In this way, pupils are not advantaged or disadvantaged by differences in difficulty of the tests over time. Test equating gives recommended cutscores, which are then considered by senior markers in the second method of standard setting used in national curriculum tests – script scrutiny.

Script scrutineers, who are senior markers, consider the work of candidates across a range of marks before coming to a judgment about the mark on which they would be prepared to say that pupils had achieved the same standard as for the previous year. Script scrutiny is a lengthy and time-consuming process, and script scrutineers need to be given a limited range of scripts to examine. The range of scripts they consider is determined by taking the cutscore recommended on the basis of test equating and then a range of mark points above and below that recommended cutscore.

In the absence of test equating evidence for level setting for the December 2007 single level tests, an alternative method of identifying a draft cutscore or small range of mark points within which the cutscore was likely to lie, needed to be developed.

General qualifications awarding bodies do not routinely use test equating for standard setting for GCSE and A level, but instead produce predicted outcomes for GCSE and A level based on statistical analysis of pupils' prior attainment. These are used to give statistically recommended cutscores on which to base script scrutiny ranges. The QCA investigated the possibility of doing a similar analysis for the December 2007 test session. It was decided that the data generated by this method would not be dependable in this context. First, there was no robust measure of prior attainment for most of the pupils sitting single level tests. Where there was data (for example, where there was key

stage 2 data for key stage 3 pupils), the time between the key stage 2 test and the single level test differed according to the year group of the key stage 3 pupils. So, even if some tentative predictions could be made about year 9 pupils on the basis of their key stage 2 results, no such predictions could be made for year 7 or year 8 pupils. It might have been possible to use predictions for year 9 pupils, to begin to establish where the cutscore might lie, but this was judged unwise because the year 9 pupils who were entered for the December 2007 single level test sessions were not necessarily representative of the whole cohort. Any predictions made on the basis of the national cohort of year 9 pupils would be unlikely to be applicable to this cohort.

The approach to identifying a mark range for script scrutiny which was used for the December 2007 test session was based on collecting additional data from all markers at the marking for the December 2007 test session. This data consisted of markers' holistic grading judgments of the scripts they had marked, an approach whose validity and reliability have support in the research literature (for example, see Wood, 1991).

Markers were asked to familiarise themselves with the APP criteria and to use them in making their judgments. A marker was asked to mark a script and then review that script and make one of five judgments:

0: that the pupil was working below the level of the test and probably should not have been entered for the test because it was too difficult for him or her.

1: that the pupil was working at just below the threshold of the level of the test.

2: that the pupil was working at just above the threshold of the level of the test.

3: that the pupil was working securely at the level of the test.

4: that the pupil had performed so well in the test that it was possible that he or she should have been entered for a higher level test.

There was a concern that markers might be predisposed to make judgments based on the marks pupils had received, so, for example, consider approximately 70 per cent of the marks to represent secure performance and 50 per cent of the marks to represent threshold achievement of the level. Such an approach would be appropriate if the test was of a known and appropriate level of difficulty, but, given that the December 2007 single level tests had not been pre-tested, this was not the case. For that reason, markers were asked to try not to make their judgments purely on the basis of the marks on a script, but to focus on the quality of the work produced. In that way, if the tests had

proved, for example, more difficult than anticipated, there was a possibility that markers might be able to compensate for this in their judgments.

Markers' judgments were tallied in relation to the total marks awarded to scripts. The modal judgment for each mark point was then considered and a range of mark points selected. The highest mark point was the point at which the modal judgment was 'secure' and the lowest mark point was that at which the modal judgment was 'just above level'. Scripts at each of these mark points were selected for presentation to senior markers at script scrutiny.

To support senior markers' judgments at script scrutiny, the widest possible range of evidence was made available. Scrutineers had single level test scripts from the December 2007 test session, not only at the range of marks identified on the basis of markers' holistic judgments, but outside that range. They could request these if they felt they would be helpful. They had national curriculum test level descriptions and the APP materials, and many scrutineers also brought notes they had made as they marked scripts.

Scrutineers were asked to begin by making completely independent judgments. Usually, in the context of national curriculum test script scrutiny, the meeting begins with a consideration of the question paper for that particular year and a discussion about the characteristics of performance scrutineers would expect to see at the various level thresholds. Given that the cutscore was to be set at a previously unknown standard of performance, scrutineers were asked to begin by making independent judgments about pupil performance. Once those judgments had been made, they could be tabulated and the magnitude of the differences between scrutineers' judgments could be considered. Despite the difficulties, there was only limited evidence of a little more variation in scrutineers' judgments than was usually the case for statutory tests and scrutineers were able, for all three subjects, to identify a mark or narrow range of marks they could recommend to the final level setting meeting.

Final level setting took place on 10 January 2008 for mathematics and on 11 January 2008 for English reading and writing. Although the December 2007 single level tests were the first round of a pilot and not statutory, as far as possible the level setting meetings followed the code of practice for statutory national curriculum assessments (QCA, 2007a). The meeting was observed by staff from the QCA's Regulation and Standards Group (now Ofqual). There were separate meetings for each subject, at which level thresholds were set for all the tests in that subject.

A range of data was presented at the level setting meeting. Although the marking data was still undergoing cleaning at that point, a substantial proportion of data was on the system, enabling the outcomes of judgments made at the meeting to be known with a high degree of certainty. Marker holistic judgment data was also available, as were the outcomes of the judgments made at the script scrutiny meetings.

It was noted that the mathematics scrutineers had arrived at a mark at which they could agree that the pupil had demonstrated secure performance at a level and then had recommended a cutscore at one or two marks above that mark, to be more confident about their judgments. In contrast, reading and writing scrutineers had recommended a cutscore at the mark on which they agreed that pupils had demonstrated secure performance. It was agreed that, to bring the mathematics judgments into line with the reading and writing judgments, the mathematics scrutineers' recommended thresholds should be reduced by one or two marks, bringing the recommendations back to their original levels. This was agreed by the mathematics senior marker.

The meeting discussed the available data and the likely outcomes of level setting, and agreed the thresholds for each test in the subject concerned, beginning with the level 5 test, moving up to the level 6 test and then down to the level 4 and level 3 tests, to follow the order of judgments required by the national curriculum test code of practice. As a result of this process, level thresholds were agreed.

Table 7 below gives details of the outcomes from the December 2007 single level test session:

Table 7: Level threshold scores and proportion of pupils who achieved the level for each subject at each level

Subject	Level	Threshold	Proportion achieving the level (overall)	Proportion achieving the level (KS3)	Proportion achieving the level (KS2)
English writing	3	19	67.2	25.0	69.1
	4	19	58.7	22.9	68.7
	5	17	39.7	37.8	54.7
	6	18	38.4	38.4	0.0
English reading	3	19	75.4	34.3	77.4
	4	23	63.5	29.1	66.8
	5	19	20.8	18.0	28.0
	6	17	4.3	4.3	0.0
Mathematics	3	27	71.2	22.1	74.1
	4	30	64.2	12.9	80.7
	5	25	14.9	8.9	49.4
	6	23	18.6	18.6	33.3

The relatively low rate of level achievement for some single level tests was a matter of concern. In the absence of data from pre-testing, it was not possible to know, with certainty, why the rates of level achievement for some tests had been lower than anticipated. The possibility that the cutscores had been set too high was considered and a modelling exercise was carried out to investigate where the cutscores would have needed to have been set to bring about a high rate of level achievement (say, 90 per cent). For almost half of the single level tests in December 2007, the cutscore would have had to be reduced by over 10 raw marks, and for four of the tests the cutscores would have needed to be set at 10 raw marks or lower. A review of the evidence available – the judgments made by the markers and by the script scrutineers - indicated that there was no evidence for lower cutscores.

Level setting for single level tests for June 2008

The intention had been to carry out standard setting for single level tests on the basis of test equating from December 2008. Given the outcomes from the December 2007 test session, it was agreed that a second pre-test would be conducted for the June 2008 tests. The requirement to appoint, at short notice, test equating agencies and a pre-test administration agency and then recruit the 35,000 pupils required for pre-test purposes meant that the pre-test for the June 2008 single level tests took place concurrently with the administration of the live tests.

In addition to practical issues, there were technical issues to address. If single level tests were to be used for reporting against the Public Service Agreement targets, they would need to be linked to national curriculum tests.

An invitation to tender was issued to test equating agencies. On the basis of the tenders received, it was decided that a number of different approaches to equating would be used, to ensure that the link had been made on the most robust evidence base possible. Two approaches to equating used Item Response Theory, with one using a one-parameter logistic model and the other a two-parameter graded response model. Two approaches to equating using classical test theory approaches were used, one using linear equating and the other using equipercentile equating. The purpose underlying each form of equating was the same: to identify equivalent scores on different tests.

A main contract and a secondary contract for test equating were awarded to two separate agencies.

One agency (the main equating agency) was commissioned to carry out equating for all subjects at all levels. A second agency (the secondary equating agency) was commissioned to carry out a particular approach to equating (live-test to live-test) for all subjects and to carry out additional equating for reading tests, where it was judged, on the basis of experience from national curriculum tests, that equating might be the most problematic and a wider evidence base might be needed.

As noted above, the pre-test for the June 2008 test session was held concurrently with the live tests and as the December 2008 single level tests were developed concurrently with the June 2008 tests, they were pre-tested concurrently with them. This meant that there were, in effect, three test sessions running during June 2008: the live test administration of the June 2008 single level tests, a concurrent pre-test 2 for the June 2008 single level tests and a pre-test 2 for the December 2008 single level tests. The equating model was designed to link these tests to each other and also to national

curriculum tests. The large number (35,000) of pupils required was to enable links to be made for different national curriculum year groups.

Strictly speaking, the work for both agencies was a linking exercise rather than an equating exercise, because the constructs being assessed were not the same across tests. Prior to the equating exercises being carried out, it was anticipated that there would be a likely variation in the cut-scores recommended, arising in part from the different approaches used and, in part, from the requirement of the test model to assess across key stages 2 and 3.

The purposes underlying the way in which the test equating model was designed were to link single level tests:

- with national curriculum end-of-key-stage tests
- within a subject, placing them on a single underlying psychometric scale for each subject
- from the June 2008 and December 2008 test sessions.

The model was large and complex (see Appendix A), not only because of the number of tests being linked, but also because the requirement for age-independence meant that links needed to be made, for each test, at each level, using pupils from a range of year groups. Although in principle pupils from year 3 through to year 9 were eligible for entry for any single level test at any level from level 3 through to level 8, it was decided not to include pupils from years 3 and 4 on the grounds that they would be very unlikely, at the time of the pre-test, to have covered the necessary curriculum. To achieve the purposes inherent in the design of the model, the pre-test consisted of the following groups of pupils:

- pupils taking a June single level test and an existing anchor test (for reading only)
- pupils taking a June single level test and a newly-constructed anchor test (all subjects)
- pupils taking two consecutive levels of June single level tests (for reading and writing only)
- pupils taking a December single level test and an existing anchor test (for reading only)

- pupils taking a December single level test and a newly-constructed anchor test (all subjects)
- pupils taking two consecutive levels of December single level tests (for reading and writing only)
- pupils taking a June and December single level test at the same level (all subjects).

'Existing anchor tests' are anchor tests that are currently in use for national curriculum test pre-testing. There are key stage 2 and key stage 3 anchor tests, and they are national curriculum tests that are kept and administered securely, year-on-year, to pupils in pre-tests. They address the same constructs as national curriculum tests, that is either the key stage 2 or key stage 3 programmes of study, and are multi-level tests.

'Newly-constructed anchor tests' were developed by the main equating agency for use in the context of single level tests. The principle underlying these new anchor tests was to avoid the issue of the anchor tests assessing different constructs by developing single level anchor tests made up from items from national curriculum tests. The development and use of these newly-constructed anchor tests is considered in more detail below.

The two equating agencies worked independently, but, once they had initial findings, these were shared and discussed.

Both agencies, in their analyses of the pre-test data, found evidence of pre-test effects, manifested in a number of ways. Average omission rates and the proportion of pupils not reaching the end of the paper (defined by the omission of the final item) were noted, as shown in Table 8 below.

Table 8: Omission rates and proportion not reached under live test and pre-test conditions

	Pre-test		Live test	
	Average percent omission rate	Percent of pupils not reaching end	Average percent omission rate	Percent of pupils not reaching end
R3	6.05	9.9%	5.03	10.7%
R4	7.22	11.1%	5.13	10.2%
R5	4.54	12.8%	4.81	14.1%
R6	8.31	23.4%	5.60	22.6%
R7	4.09	14.0%	2.02	9.4%
M3 paper 1	2.60	12.1%	1.70	8.7%
M3 paper 2	2.75	9.7%	1.77	8.1%
M4 paper 1	3.84	14.4%	3.33	12.0%
M4 paper 2	5.35	15.6%	4.12	12.1%
M5 paper 1	5.84	30.7%	5.20	23.6%
M5 paper 2	6.22	19.3%	5.26	13.1%
M6 paper 1	5.91	20.1%	3.51	11.8%
M6 paper 2	6.43	25.7%	3.17	13.7%
M7 paper 1	5.25	13.7%	3.26	6.0%
M7 paper 2	7.27	10.1%	3.84	5.5%
M8 paper 1	3.77	9.4%	1.52	0%
M8 paper 2	9.14	8.3%	3.16	0%

The figures indicate a decrease in average omission rates from pre-test to live test administration for all tests, with the exception of reading level 5, supporting the existence of a pre-test effect. For mathematics, the proportion of pupils not reaching the end of the test is higher in a pre-test situation.

In terms of mean scores on the tests, there were differences between the pre-tests and the live tests. It would not be appropriate to make direct comparisons between the mean scores for pupils in the pre-test and in the live test. Pupils for the pre-test were carefully selected so that there was control over relevant factors (such as ability), whereas there

was less control over pupils who were entered for the live test. Pupils from each of the groups would, therefore, be unlikely to be similar. If relevant factors were taken into account, comparisons would be possible, however. This could be done by constructing a regression model. The linear regression coefficients from such a model are given in Table 9 below.

Table 9: Significant linear regression coefficients for total test score

	Reading		Writing		Maths	
Level 3	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Stakes	1.42	0.00	1.43	0.00	1.12	0.00
TA* level	2.26	0.00	1.46	0.00	3.10	0.00
Gender	0.34	0.02	1.11	0.00		0.97
National curriculum year	-0.69	0.00	-0.84	0.00	-1.12	0.00
% variance explained	19.9%		20.4%		17.3%	
N	5859		5329		5173	
Level 4	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Stakes	1.17	0.00	0.62	0.00	0.93	0.00
TA level	2.19	0.00	1.20	0.00	2.84	0.00
Gender	-0.26	0.05	1.37	0.00		0.33
National curriculum year	-0.85	0.00	-1.19	0.00	-3.11	0.00
% variance explained	25.4%		26.0%		34.4%	
N	6770		4602		4226	

	Reading		Writing		Maths	
Level 5	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Stakes		0.20		0.72	0.65	0.04
TA level	1.85	0.00	1.21	0.00	2.80	0.00
Gender	0.57	0.00	0.57	0.00	-0.56	0.03
National curriculum year	-0.67	0.00	-0.78	0.00	-3.88	0.00
% variance explained	21.3%		21.3%		29.6%	
N	3889		3156		3243	
Level 6	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Stakes		0.09	1.77	0.00	2.40	0.00
TA level	1.64	0.00	1.18	0.00	2.24	0.00
Gender		0.95	0.51	0.04	-0.77	0.03
National curriculum year	-0.85	0.00		0.44	-2.76	0.00
% variance explained	23.6%		27.8%		18.9%	
N	2236		1370		2060	
Level 7	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Stakes	3.32	0.00	7.20	0.00		0.35
TA level	0.73	0.00	1.62	0.00	1.55	0.04
Gender		0.84		0.48		0.68
National curriculum year	1.04	0.00	-1.45	0.01	3.44	0.05
% variance explained	12.7%		31.9%		10.9%	
N	1135		447		865	

* Teacher assessment

Issues around differential item functioning (DIF) were noted by both agencies. A differential item functioning (DIF) analysis was carried out on items from the pre-test and live tests. This indicated that a relatively large number displayed DIF, though the majority were categorised as negligible in terms of severity. Overall, a greater proportion of reading items exhibited DIF than mathematics items. In its analysis, the main equating agency gave a detailed technical consideration to dealing with DIF. In essence, common items were used to link the two ability scales by carrying out separate analyses to estimate the parameters on the national curriculum test items using live data and on the single level test and newly-constructed anchor test items and then estimating a 'shift parameter' to bring all the items onto the same scale.

Issues around the tests being linked

As noted above, a key assumption underlying test equating is that the tests being equated assess the same construct and it was not clear that this requirement was met, in this instance.

For the mathematics tests, there were two issues to deal with. First, national curriculum tests included a separate mental mathematics test but single level tests did not. Second, national curriculum tests are tiered, whereas single level tests are not. The agencies worked on the basis that, for national curriculum tests, levels across tiers were equivalent. For the reading tests, national curriculum tests include the assessment of Shakespeare, whereas single level tests do not, so the contribution of the Shakespeare element to the test score on a national curriculum test had to be addressed.

Equating using item-response theory and newly-constructed anchor tests

The main equating agency was contracted to equate for all subjects at all levels. It advised that item-response theory (IRT) was the most appropriate methodology for the mathematics and reading tests, and the approach it used was the one-parameter logistic model. It also advised that the use of IRT was inappropriate for these writing tests and that a non-IRT approach was required.

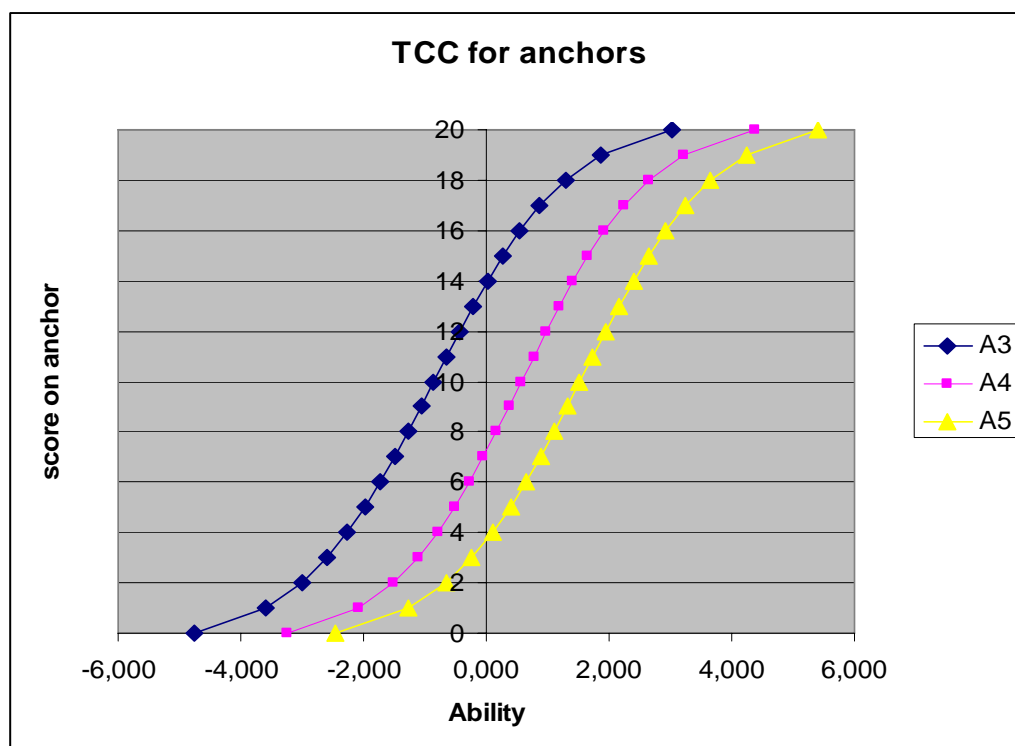
The principle underlying the approach of the main equating agency was to link single level tests with national curriculum tests by developing newly-constructed anchors derived from national curriculum tests. The rationale behind this approach was to minimise issues associated with linking a single level test with a multi-level national curriculum test or multi-level national curriculum test anchor, by linking the single level test to an anchor test consisting of an appropriate subset of national curriculum test items.

The selection of items for the newly-constructed anchor tests was based on optimisation of information in the relevant region on the ability axis, which meant that items were selected which were optimal in terms of measurement around the cutscore at a particular level on the national curriculum test. The items were identified by taking data from previous administrations of national curriculum tests for each subject (reading, writing, mathematics) from a number of years (2003 to 2007). The data was analysed using an IRT model, beginning with the calibration of each item set using the Rasch model. Score ranges on each key stage test were translated into an ability value, with the centres of the ability ranges corresponding to the observed score ranges. An algorithm was then used to optimise the information over this value, under constraints on the composition of the anchor tests. The major constraint was that the anchors had to be delivered in no more time than a single level test. In practice, this meant that the tests should have durations of no longer than 50 minutes.

There were a number of constraints on the development of the newly-constructed anchor tests, such as the need to exclude mental mathematics test items and the requirement for some items to be kept together in groups (for example, questions relating to particular texts for the reading anchor tests had to be kept together). With these constraints in place, a set of anchor tests was produced. For key stage 2, there was a level 3/4 and a level 4/5 anchor for which the level 4 items were common. For key stage 3, there was a level 4/5 anchor, a level 5/6 anchor and a level 6/7 anchor. It was not possible to construct common anchors for key stage 2 and key stage 3 because there were no common items across the national curriculum tests for these two key stages.

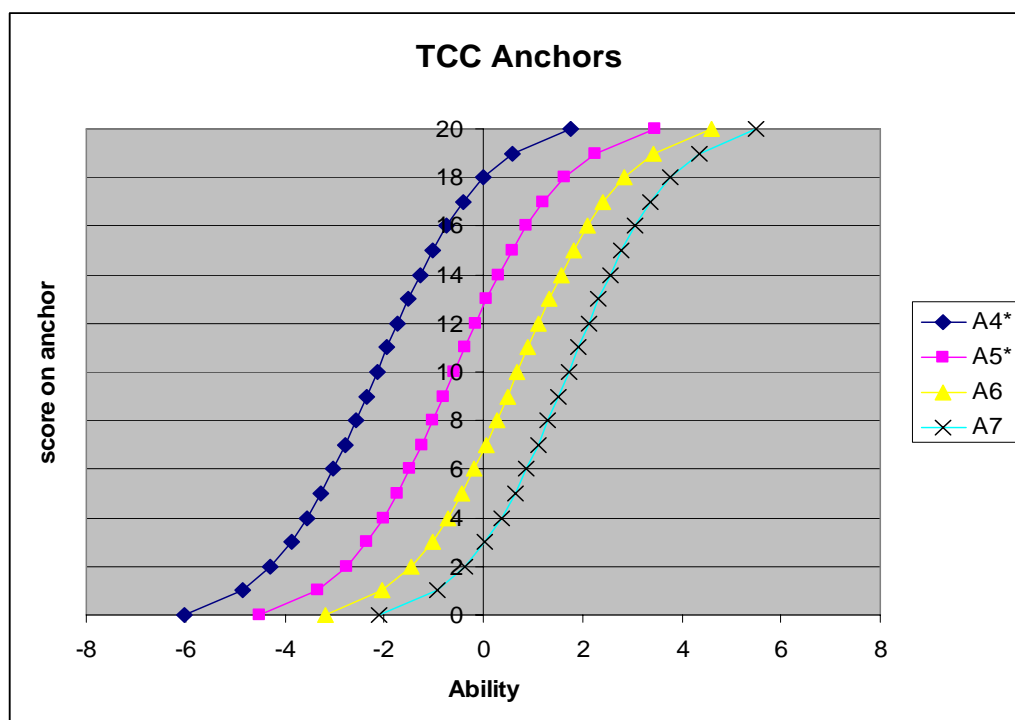
A consideration of the anchor tests for mathematics shows clearly how the approach worked. Figures 1 and 2 below show the test characteristic curves for the mathematics anchor tests. They show that the anchor tests consisted of subsets of items of similar difficulty, with anchor tests at higher levels being more difficult than those at lower levels.

Figure 1: Test characteristic curves for the key stage 2 anchor tests (mathematics)



Key: A3 (anchor test, L3 SLT), A4 (anchor test, L4 SLT), A5 (anchor test, L5 SLT)

Figure 2: test characteristic curves for the key stage 3 anchor tests (mathematics)



Key: A4 (anchor test L4 KS3), A5 (anchor test L5 KS3), A6 (anchor test L6), A7 (anchor test L7)

Achieved-score equating, using newly-constructed anchor tests

The approach developed by the main equating agency for standard setting for single level tests in writing was achieved-score equating. This involves the cross-tabulation of:

A: the score on the newly-constructed anchor test

R: the score on the remaining part of the national curriculum test and

S: the score on the single level test.

Different cross-tabulations can be produced depending on the equating design, and for the 2007 live test data the main equating agency obtained a cross-tabulation of the score on the anchor test (A) and the score on the remaining part of the national curriculum test (R). This provided information on the marginal distribution of R-scores and A-scores. From the pre-test, a cross-tabulation of scores on the anchor test (A) and the single level test (S) was produced. Together, these cross-tabulations contained all available information for making inferences about $(A, R)|S$ (the joint distribution of A and R conditional upon S). Using that, the main equating agency estimated the percentage of pupils with a single level test score who would achieve the level in the national curriculum test, using the national curriculum test cutscores and determining the cutscore on the single level test that would correspond to a 50 per cent probability of the pupil achieving the level in the national curriculum test.

Live-test to live-test equating

The rationale behind this approach was to link national curriculum tests with single level tests by using pupils who had sat a single level test and a national curriculum test as a link. This approach to equating, therefore, used pupils in years 6 and 9 and was carried out for reading, writing and mathematics tests.

The strength of this approach was that it enabled a direct link to be made, in a live-test situation, between pupils taking live national curriculum tests and live single level tests. Theoretically, this should avoid a pre-test effect, but only if motivational and preparation effects are the same in both test situations.

For reading and mathematics, the secondary equating agency analysed the data using two-parameter IRT true score equating. For writing, where the secondary equating agency recommended that IRT was not the most appropriate method, they used equipercentile equating.

Equating using anchor tests

The secondary equating agency also carried out pre-test equating for the reading tests, using anchor tests. This analysis used the existing anchor tests for key stage 2 and key stage 3 and the newly-constructed anchors. The existing anchor tests are well established. The secondary equating agency analysed the performance of the newly-constructed anchor tests, stating that they had performed very well overall, with low rates of questions being omitted or not reached and only a few items achieving lower than desired discrimination characteristics.

In its analysis, the secondary equating agency considered the fullest possible range of equating designs, using:

- the newly-constructed and the existing anchor tests in the same analysis across all pupils
- the newly-constructed and the existing anchor tests in the same analysis with pupils from key stage 2 and then, separately, pupils from key stage 3
- the existing anchor tests across all pupils
- the existing anchor tests with pupils from either key stage 2 or key stage 3 separately.

In order to ensure sufficient linking between tests to produce adequate data to allow the QCA to place confidence in the outcomes, a large and complex model, involving approximately 35,000 participating pupils, was required. Pupils from both key stages 2 and 3 had to take combinations of papers from the 36 single level tests (for June and December 2008) and national curriculum test anchors (either existing anchors or newly-constructed ones). The model design and thus the magnitude of the pre-test were largely a function of the nature of single level tests and, in particular, of the requirement for age-independence. It was also a function of the need, for the first attempt at linking, to generate the best possible evidence base for evaluating the tests and the test model.

Pupils were recruited who had been judged, by their teachers, to be working at approximately the level of the tests they were taking. While it was not desirable, for example, for a pupil working at level 3 to be entered for a level 7 test, it was desirable for there to be pupils who might be working just below and just above the level of the test to participate. For this reason, teacher assessment was judged an adequate measure, as it would be likely to include some pupils who would perform at a higher level in a test and some who would perform at a lower level. In order to avoid group effects in the data,

attempts were made to ensure that no more than 15 pupils from a school were in a single group.

Initial findings from test equating for the June 2008 single level tests

Both pre-test and live-test data were available for analysis. The analysis for single level tests in mathematics, carried out by the main equating agency, using an IRT approach, gave the recommended cutscores set out in Table 10 below.

Table 10: Recommended cutscores for single level tests in mathematics

Test	Total marks	Recommended cutscore
Level 3	50	30
Level 4	50	28
Level 5	48	30
Level 6	47	18–20
Level 7	47	18
Level 8	47	17

In the context of its analysis, the main equating agency pointed out that the single level test model was predicated on the notion that levels 4 and 5 in particular (but, by inference, all levels) were the same in terms of standards for both key stage 2 and key stage 3, but the analysis indicated that this was not the case. To ensure model fit, the agency equated levels 3, 4 and 5 using the key stage 2 standard and levels 6, 7 and 8 using the key stage 3 standard. Although it was practically possible to equate levels 4 and 5 to the key stage 3 standard, it was not technically defensible.

The analysis of the secondary equating agency was carried out independently, but painted a very similar picture. The secondary equating agency used live-test to live-test equating. Table 11 below shows the findings from the equating analysis carried out for year 6 pupils, using performance on the key stage 2 national curriculum test in mathematics and single level tests at levels 3, 4 and 5.

Table 11: Recommended cutscores from live-test to live-test equating for key stage 2

Single level test	Total marks	Compared with KS2 national curriculum tests	
		Equipercetile equating	IRT equating
Level 3	50	23	23
Level 4	50	24	25
Level 5	48	29	30

The secondary equating agency carried out a similar analysis for key stage 3. However, given the tiering structure for key stage 3, the numbers in some cases were judged to be too small for robust analysis.

Table 12: Recommended cutscores from live-test to live-test equating for key stage 3

	IRT – Tier 3–5	IRT – Tier 4–6	IRT – Tier 5–7	IRT – Tier 6–8
Level 4	23	21–22		
Level 5	21	17	23–24	
Level 6		15–16	26	20
Level 7			-	29

Bringing together the evidence, there was variation in the recommended cutscores and Table 13 below summarises the cutscore ranges:

Table 13: Summary of equating cutscore ranges

Single level test	Total marks	Cutscore mark range (KS2)	Cutscore mark range (KS3)
Level 3	50	23–30	
Level 4	50	24–28	21–23
Level 5	48	29–30	17–24
Level 6	47		15–26
Level 7	47		18–29
Level 8	47		17

Further analysis indicated that younger pupils, on average, did better than older pupils on single level tests levels 3, 4 and 5. This can be seen from Table 14 below.

Table 14: Average total mark on single level tests in mathematics, by age

Level 3 single level test			Level 4 single level test			Level 5 single level test		
Year	Average score	n	Year	Average score	n	Year	Average score	n
3	32.66	141						
4	36.48	1494	4	37.73	283			
5	38.38	2401	5	38.56	1419	5	37.03	76
6	36.37	343	6	36.09	776	6	35.69	402
7	28.35	81	7	26.29	177	7	26.48	230
8	29.07	107	8	27.60	412	8	26.57	1074
			9	29.02	180	9	26.13	485

The findings for reading

As for mathematics, analyses were carried out separately by the two agencies. The analysis carried out by the main equating agency indicated that model fit was not an issue for reading.

As with mathematics, a range of recommended cutscores came from the different analyses, but the underlying picture was coherent. It should be noted that the magnitude of the differences in recommended cutscores for key stages 2 and 3 is less for reading than for mathematics, but that is largely because the total number of marks available for the mathematics tests (approximately 50 per test) was higher than the total number of marks available for the reading tests (approximately 35 per test).

The analysis of the main equating agency indicated a range of cutscores, as given in Table 15 below.

Table 15: Recommended cutscores for single level tests in reading based on the one-parameter logistic (IRT) model

Test	Total marks	Recommended cutscore
Level 3	38	16
Level 4	37	12
Level 5	39	15
Level 6	34	10
Level 7	25	17–18

As for mathematics, the main equating agency argued that it was not technically defensible to link levels 4 and 5 to key stage 3 and, for level 4, it proved impossible. It was asked to carry out the analysis for level 5 for experimental purposes. The recommended cutscore for level 5, using key stage 3 pupils, was 17–18.

The secondary equating agency had been commissioned to analyse the reading data not only using live-test to live-test equating, but also using anchor tests to carry out analyses based on pre-test equipercetile equating and IRT modelling. As would be expected, this analysis generated a range of recommended cutscores. These are set out below.

Table 16: Recommended cutscores for single level tests in reading using live-test to live-test equating

		Compared with KS2 NCT		Compared with KS3 NCT	
		Equipercntile	IRT	Equipercntile	IRT
Level 3	Threshold	15	15	N/A	N/A
	N	271		29	
Level 4	Threshold	11	13	17	15
	N	796		91	
Level 5	Threshold	15	17	14	15
	N	484		425	
Level 6	Threshold	N/A	N/A	8	9
	N	25		310	
Level 7	Threshold	N/A	N/A	N/A	N/A
	N	0		47	

Table 17: Recommended cutscores for single level tests in reading using pre-test equipercntile equating against anchor tests

		KS2 anchor	KS3 anchor
Level 3	Threshold	16/17	
	N	535	
Level 4	Threshold	12/13	
	N	762	
Level 5	Threshold	15/17	21
	N	414	367
Level 6	Threshold		
	N		
Level 7	Threshold		
	N		

Table 18: Recommended cutscores for single level tests in reading using IRT modelling, including the newly-constructed anchors in the model

		All pupils		KS2 pupils only		KS3 pupils only	
		KS2 threshold	KS3 threshold	KS2 threshold	KS3 threshold	KS2 threshold	KS3 threshold
Level 3	Threshold	16	N/A	17	N/A	N/A	N/A
Level 4	Threshold	13–14	14	13	N/A	N/A	19
Level 5	Threshold	19	18	18	N/A	N/A	20
Level 6	Threshold	N/A	17	N/A	N/A	N/A	16
Level 7	Threshold	N/A	-	N/A	N/A	N/A	-

Table 19: Recommended cutscores for single level tests in reading using IRT modelling, without the newly-constructed anchors in the model

		All pupils		KS2 pupils only		KS3 pupils only	
		KS2 threshold	KS3 threshold	KS2 threshold	KS3 threshold	KS2 threshold	KS3 threshold
Level 3	Threshold	16	N/A	16	N/A	N/A	N/A
Level 4	Threshold	14	14	13	N/A	N/A	19
Level 5	Threshold	20	18	18	N/A	N/A	20
Level 6	Threshold	N/A	17	N/A	N/A	N/A	15
Level 7	Threshold	N/A	19	N/A	N/A	N/A	-

For the reading tests, it was technically possible to put key stage 2 and key stage 3 pupils onto the same underlying scale. The recommended cutscores for key stage 2 and key stage 3 were, however, different. While some differences might be expected, if the standards for key stages 2 and 3 were the same (or very similar), these differences would be small.

For level 4, the different cutscores for key stages 2 and 3 can be seen in Table 23 below.

Table 20: Cutscores for level 4 reading

Method	Key stage 2 cutscore	Key stage 3 cutscore
Live-test to live-test equipercentile	11	22
Live-test to live-test two-parameter IRT	11	14
Pre-test using anchor tests	12/13	n/a
Two-parameter IRT with newly-constructed anchors	13	19
Two-parameter IRT without newly-constructed anchors	13	19

For level 5, the different cutscores for key stages 2 and 3 can be seen in Table 21 below.

Table 21: Cutscores for level 5 reading

Method	Key stage 2 cutscore	Key stage 3 cutscore
Live-test to live-test equipercentile	15	21
Live-test to live-test two-parameter IRT	17	15
Pre-test using anchor tests	15–17	21
Two-parameter IRT with newly-constructed anchors	18	20
Two-parameter IRT without newly-constructed anchors	18	20
One parameter IRT with newly-constructed anchors	15	17–18

The findings for writing

As for mathematics and reading, the equating analyses for writing were carried out independently by the two equating agencies and the cutscores recommended for key stage 2 and key stage 3 pupils are given in Table 22 below.

Table 22: Cutscores for writing (achieved-score equating)

Single level test	Total marks	Cutscore mark range (KS2)	Cutscore mark range (KS3)
Level 3	27	5–12	-
Level 4	27	14–15	-
Level 5	23	20–21	11–12
Level 6	23	11–12	-
Level 7	23	16–17	-

The recommended cutscores from the equipercntile live-test to live-test equating carried out by the secondary equating agency are given in Table 23.

Table 23: Cutscores for writing (equipercntile equating)

Single level test	Total marks	Cutscore mark range (KS2)	Cutscore mark range (KS3)
Level 3	27	12	
Level 4	27	19	15
Level 5	23	17	9
Level 6	23	-	11
Level 7	23	-	-

Here, the recommendations using the key stage 3 pupils were lower than those using the key stage 2 pupils.

Level setting for the June 2008 single level tests

The draft level setting meeting was held on 15 September 2008. It was attended by representatives from the QCA and the test equating agencies and was observed by Ofqual. As for level setting for the December 2007 test session, as far as possible, the meeting complied with the requirements of the code of practice. The purpose of the meeting was to set draft cutscores and so identify a range of scripts for consideration by senior markers at script scrutiny.

The main issue was the evidence generated by the test equating agencies and a number of options were considered. The possibility of using the range of recommended cutscores to set wide script scrutiny ranges and then asking script scrutineers to make a judgment about the cutscores to set the standard was considered. However, on the basis of what was known about professional qualitative judgment (Baird & Dhillon, 2006), it would not be appropriate to expect script scrutiny to identify a standard, so this option was discounted.

The possibility of setting cutscores for key stage 2 pupils and different cutscores for key stage 3 pupils and reporting separately for each key stage was considered briefly, but, although technically defensible, it was not considered to be a plausible solution. The possibility of setting level cutscores for single level tests at levels 3, 4 and 5 using the key stage 2 recommended cutscores and for single level tests at levels 6, 7 and 8 using the key stage 3 recommended cutscores was considered to be more plausible. This would, however require a decision about whether levels 6, 7 and 8 should be reported for key stage 2 pupils and levels 3, 4 and 5 for key stage 3 pupils.

The meeting reconsidered its remit, which had been to set levels for a single level test model which was based on one set of tests which were common across both key stages. It agreed, unanimously, that, given the potential uses of the results from the test session, setting levels for pupils at both key stage 2 and key stage 3, using the given model, would not be technically defensible and so advised against doing so.

QCA then generated cutscores in readiness to respond promptly to a policy-level decision about whether cutscores should be set and, if so, on what basis this should be done.

For mathematics, IRT analyses were prioritised over the findings from the live-test to live-test equating for mathematics. The cutscores arising from this approach are given in Tables 24 to 27 below, first using the key stage 2 standard and then using the key stage 3 standard.

Table 24: Recommended cutscores using the key stage 2 standard

Single level test	IRT recommended cutscore	Live-test to live-test equating recommended cutscore	Agreed cutscore
Level 3	30	23	30
Level 4	28	24/25	28
Level 5	30	29/30	30

Table 25: Outcomes using the key stage 2 standard

Single level test	% of key stage 2 pupils achieving the level	% of key stage 3 pupils achieving the level
Level 3	89.1	51.2
Level 4	92	52.9
Level 5	82	34.9

Table 26: Recommended cutscores using the key stage 3 standard

Single level test	IRT recommended cutscore	Live-test to live-test equating recommended cutscore	Agreed cutscore
Level 3	No information available: impossible to set a cutscore		
Level 4	No information available	28 (n=167)	Impossible to set a single cutscore, but >20
Level 5	20*	17-23	20
Level 6	18–20 (18)**	15–26	18
Level 7	18	29–45	18
Level 8	17	No information available	17

* Model does not fit for level 5.

** Where a range of cutscores is recommended, the principle was to take the lowest.

Table 27: Outcomes using the key stage 3 standard

Single level test	% of key stage 2 pupils achieving the level	% of key stage 3 pupils achieving the level
Level 3	N/A	
Level 4	N/A	
Level 5	99	83.1
Level 6	100	92.9
Level 7	-	92
Level 8	-*	100

* One pupil from KS2.

For reading, as for mathematics, IRT analyses were prioritised over other approaches to equating. Both equating agencies had carried out IRT-based equating, with one using

the two-parameter model currently used in national curriculum tests and the other using the one-parameter logistic model it had recommended as the most appropriate for single level tests. It was agreed that priority should be given to the approach recommended by the main equating agency, so priority was given to the recommendations from the one-parameter logistic model. These were not substantially different from the recommendations from the two-parameter IRT model used by the secondary equating agency.

Table 28: Recommended cutscores using the key stage 2 standard

Single level test	IRT (one-parameter) recommended cutscore	Agreed cutscore
Level 3	16	16
Level 4	12	12
Level 5	15	15

Table 29: Outcomes using the key stage 2 standard

Single level test	% of key stage 2 pupils achieving the level	% of key stage 3 pupils achieving the level
Level 3	96.4	74.6
Level 4	97.9	81.5
Level 5	88	73.9

Table 30: Recommended cutscores using the key stage 3 standard

Single level test	IRT (one–parameter) recommended cutscore	Agreed cutscore
Level 4	Not available*	Not available*
Level 5	17	17
Level 6	10	10
Level 7	17	17

* The two-parameter IRT model recommended a threshold of 19 marks here.

It should be noted that it was not possible to set a level 3 recommended cutscore based on the key stage 3 standard since level 3 is not awarded in key stage 3 English reading tests.

Table 31: Outcomes using the key stage 3 standard

Single level test	% of key stage 2 pupils achieving the level	% of key stage 3 pupils achieving the level
Level 3	N/A	
Level 4	76.5	41.9*
Level 5	77.6	60.5
Level 6	63.5	
Level 7	32.1	

* Using the two-parameter IRT model.

The issues around equating for writing were considered at length. The two agencies had agreed that IRT was not an appropriate approach for these tests. One agency had used a live-test to live-test equipercentile equating approach and the other had used an achieved-score equating approach. For consistency, the live-test to live-test equipercentile equating approach was selected and applied across all tests. For level 7, there was no live-test to live-test equipercentile equating recommended cutscore and

evidence therefore came from the achieved-score equating for that level. The agreed cutscore here was 16.

Table 32: Recommended cutscores using the key stage 2 standard

Single level test	Live-test to live-test equipercentile recommended cutscore	Agreed cutscore
Level 3	12	12
Level 4	19	19
Level 5	17	17

Table 33: Outcomes using the key stage 2 standard

Single level test	% of key stage 2 pupils achieving the level	% of key stage 3 pupils achieving the level
Level 3	99.3	81.0
Level 4	92	60.9
Level 5	61.1	39.2

Table 34: Using the key stage 3 standard

Single level test	Live-test to live-test equipercentile recommended cutscore	Agreed cutscore
Level 3	N/A	
Level 4	15	15
Level 5	9	9
Level 6	11	11
Level 7	16*	16

* From achieved-score equating.

Table 35: Outcomes using the key stage 3 standard

Single level test	% of key stage 2 pupils achieving the level	% of key stage 3 pupils achieving the level
Level 3	N/A	
Level 4	98	83.6
Level 5	98.6	91.8
Level 6	50*	71.9
Level 7	-	57.1

* Only 20 key stage 2 pupils entered for this test.

Levels were set for key stage 2 pupils, using the key stage 2 standard, at levels 3, 4 and 5. Levels were not set for key stage 3 pupils and single level tests were discontinued at key stage 3 from the Making Good Progress pilot.

Results from the June 2008 single level tests session

The results from the June 2008 single level tests session comprised levels for pupils at key stage 2, covering levels 3, 4 and 5. Outcomes, in terms of the percentage of key stage 2 pupils who were entered for the test and achieved the level, are given below.

Table 36: Mathematics

Single level test	Cutscore	% of key stage 2 pupils achieving this level
Level 3	30	89.1
Level 4	28	92.0
Level 5	30	82.0

Table 37: Reading

Single level test	Cutscore	% of key stage 2 pupils achieving this level
Level 3	16	96.4
Level 4	12	97.9
Level 5	15	88.0

Table 38: Writing

Single level test	Cutscore	% of key stage 2 pupils achieving this level
Level 3	12	99.3
Level 4	19	92.0
Level 5	17	61.1

Interpreting the findings from the June 2008 test session

The findings from the analysis of the pre-test data gathered for the June 2008/December 2008 single level tests need to be interpreted with caution. In the case of mathematics, whilst it was clear that the model did not fit for key stage 3 pupils, the information available at the time could not confirm the reasons for this. The removal of statutory and pilot testing at key stage 3 meant that, for practical reasons, undertaking a full-scale investigation was not justifiable.

Turning to the differences in cutscores for levels that cover both key stages (primarily levels 4 and 5), the cutscores for key stage 2 pupils were higher than those for key stage 3 pupils. It does not follow, however, that standards at levels 4 and 5 were lower for key stage 3 than for key stage 2. It is equally legitimate to infer that the standards are different. A cross-key stage comparability study (Mason 2003) compared the performance of pupils from years 6 and 9 on questions taken from the key stage 2 and key stage 3 mathematics tests and indicated qualitative differences in the nature of performance for pupils in years 6 and 9, with the performance of year 6 pupils being such that they scored higher marks than year 9 pupils.

For reading, it was *technically* possible to put the key stage 2 and key stage 3 pupils onto the same scale, but the recommended cutscores for key stages 2 and 3 were different, with higher cutscores recommended for key stage 3 than key stage 2. Again, it might be inferred that the standard for key stage 3 is higher than that for key stage 2, but, again, they may simply have been different.

A series of small-scale studies was carried out by the NFER/UCLES, the aim of which was to investigate the possibility of developing reading tests that could be used across key stages (NFER/UCLES, 1998), so the overarching aim of the project was very similar to the aim for single level tests. The most significant finding, from the reading study, was that pupils who were awarded the same level at different key stages 2 and 3 performed *differently* on the core of common questions, with pupils at the higher key stages tending to obtain higher scores (NFER/UCLES, 1998, p10). The report generated a number of hypotheses to begin to explain these differences:

It may be that the common questions were testing knowledge of the world or thinking skills which tend to increase with age. Alternatively, it may be that they were testing skills in English that reflect the programmes of study in such a way that, for example, pupils who had followed the Key Stage 2 or 3 curriculum have acquired some of these skills even though their level of performance was still at Level 3. (NFER/UCLES, 1998, p10.)

The evidence for writing indicated that the cutscores for key stage 3 pupils should be lower than those for key stage 2 pupils, indicating that standards were different. Differences were also noted in a cross key stage comparability study carried out by Green et al (2003). This study found that pupils from key stage 2 who had scored levels 4 or 5 in national curriculum tests achieved lower scores in the study than their key stage 3 counterparts. Comparisons of matched scripts from pupils at key stages 2 and 3 showed that, at level 5, the writing of key stage 3 pupils was qualitatively different from that of key stage 2 pupils – in general, it was judged to be more sophisticated.

A very small-scale investigation of the differences in performance between pupils entered for the live single level tests for June 2008, carried out with a small group of senior markers from both key stage 2 and key stage 3 backgrounds, supported the notion that standards are different. In their professional judgment, key stage 2 pupils' writing was 'tidier' than that of key stage 3 pupils. They had more control of their sentence structure and their punctuation and grammar tended to be correct. They seemed to be more in control of their writing and were more confident. Key stage 3 pupils wrote more and where they responded well to a task, they demonstrated more extensive thinking and

creativity. However, many of them seemed overwhelmed by the task. They also made technical mistakes, with one marker suggesting that they had been taught to focus on engaging the reader and producing an effect, so they had lost sight of the technical detail and could not gain credit, given the mark schemes.

The December 2008 test session

As noted above, the Secretary of State announced on 14th October 2008 that the single level test pilot would no longer continue at key stage 3. By that point, the December 2008 tests, covering levels 3, 4, 5 and 6 had been developed and pre-tested. This meant that the December 2008 tests were based on the previous test model. There was not enough time to develop a new suite of tests for December 2008, so both equating agencies were asked to take account, in their analyses, of the fact that only key stage 2 pupils were eligible for test entry. They made sure that their analysis used only key stage 2 pupils, and checked carefully to ensure that there were no issues arising because of the reduction in the number of pupils who could be used in the analysis. Although it was too late to make changes to the test papers, the test development agencies were asked to review their mark schemes to ensure that, as far as possible, key stage 2 pupils could be given credit for demonstrating characteristics of key stage 2 performance on the December 2008 tests. Senior markers were asked to make markers aware of how the mark schemes should be interpreted for evaluating key stage 2 pupils' work.

For the December 2008 tests, the main equating agency carried out the analysis for the reading and mathematics tests and the secondary agency carried out the equating for the writing tests. The approaches taken for these tests mirrored those taken for June 2008.

For single level tests in writing, the national curriculum anchor tests were used to link single level tests to national curriculum tests. Two equating designs were used: a direct equate, from the June pre-test to the December pre-test (mirroring what was done for the June 2008 test session) and a chained equate, from the June test and the anchor to the December test and the anchor. For both designs, both linear and equipercentile equating were carried out. Additionally, equipercentile equating with loglinear smoothing was carried out, for verification.

For level 3 and level 4 tests, the outcomes from linear and equipercentile equating aligned. For the level 5 test, the direct equating indicated a cutscore of 16 and the chained equating indicated a cutscore of 18. It was agreed that, as the direct equating minimised the potential accumulation of error, this was preferred.

Script scrutiny had not been required to set standards for the June 2008 single level tests, but was used as part of the level setting process for the December 2008 tests. Script scrutineers were asked to use their professional judgment to verify the cutscores indicated by the test equating. The senior marker for each of levels 3, 4 and 5 (and also level 6 in the case of mathematics) worked together to review scripts from each single level test. They were given scripts at three mark points: the recommended cutscore, the mark point above it and the mark point below it. Three scripts were available to each scrutineer on each mark point, and they were asked whether, in their professional judgment, scripts on the recommended cutscore and the mark points immediately adjacent to it would be acceptable as evidence of attainment at the level of the test. For mathematics, senior markers agreed unanimously that scripts on the cutscores recommended by equating were acceptable as evidence that pupils had demonstrated attainment at the threshold of the level of the test. The three scripts from pupils entered for the level 6 test (which represented the whole of the entry for the test at this level) were not reviewed, as they were all on mark points substantially above any possible cutscore.

For reading and for writing, senior markers agreed unanimously that scripts on the cutscores recommended by equating were acceptable as evidence that pupils had demonstrated attainment at the threshold of the level of the test.

Final level setting

The final level setting meeting for single level tests for December 2008 was held on 14 January 2009.

Mathematics

The cutscores recommended by equating were validated by senior markers and are shown, along with the outcomes, in Table 39 below.

Table 39: Final outcomes: single level tests in mathematics

Level	Cutscore (June 08)	Proportion of pupils achieving the level in the test (June 08)	Cutscore recommended by test equating	Cutscore validated by senior markers	Agreed cutscore (Dec 08)	Proportion of pupils achieving the level in the test (Dec 08)
3	30/50	89.1	27	27	27/50	80.2
4	28/50	92	26	26	26/50	89.8
5	30/48	82	30	30	30/50	91.8

For the level 6 tests, in the absence of equating evidence, a notional cutscore of 25/50 marks was set, reflecting the location of the cutscore for the other level tests. It was agreed that, if there were substantial rises in test entries for level 6 in mathematics, and scripts around the probable cutscore became available, the judgment about the notional cutscore would need to be revisited.

Reading

The cutscores recommended by equating were validated by senior markers and are shown, along with the outcomes, in Table 40 below.

Table 40: Final outcomes: single level tests in reading

Level	Cutscore (June 08)	Proportion of pupils achieving the level in the test (June 08)	Cutscore recommended by test equating	Cutscore validated by senior markers	Cutscore (Dec 08)	Proportion of pupils achieving the level in the test (Dec 08)
3	16/38	74.6	16	16	16/36	92.1
4	12/37	81.5	10	10	10/34	96.1
5	15/39	73.9	12	12	12/36	87.0

Writing

The cutscores recommended by equating were validated by senior markers and are shown, along with the outcomes, in Table 41 below.

Table 41: Final outcomes: single level tests in writing

Level	Cutscore (June 08)	Proportion of pupils achieving the level in the test (June 08)	Cutscore recommended by test equating	Cutscore validated by senior markers	Cutscore (Dec 08)	Proportion of pupils achieving the level in the test (Dec 08)
3	12/27	99.3	13	13	13/27	97.4
4	19/27	92	17	17	17/27	87.3
5	17/23	61.1	16	16	16/23	54.4

Taking forward the development of single level tests

On the basis of the evidence generated by the pilot and supporting research, the test model being taken forward is that there will be four single level tests available – a level 3 test, a level 4 test, a level 5 test and a level 6 test - for each of English reading, English writing and mathematics.

The primary purpose of the tests is to confirm a teacher assessment that a pupil is working at a particular national curriculum level. A test confirms a teacher assessment by providing a parallel source of evidence and an independent assessment of a pupil's attainment at a level. Levels will be awarded (or not) solely on the basis of test results.

To ensure that national standards can continue to be measured consistently, single level tests must carry forward the standard of current end-of-key-stage tests, so the standard for single level tests is to be the key stage 2 standard. The current level 6 standard is based on the key stage 3 programme of study, so this means that a new standard will need to be set for level 6, based on the key stage 3 programme of study and reflecting the kinds of performance to be expected of high attaining pupils at key stage 2.

The standards for current end-of-key-stage tests are set using the performance of year 6 pupils, on a test sampling from the whole of the key stage 2 programme of study and marked using mark schemes developed to reflect expectations about the performance of these particular pupils. This means that, for each single level test, the construct being

assessed will need to be the whole of the key stage 2 programme of study, so each test will need to sample appropriately from that programme of study. The intention here is to ensure that standards are carried forward and also to ensure breadth of teaching and learning.

Single level tests are age-independent, because a teacher can enter a pupil for a single level test at any point from year 3 to year 6, once she or he has made the judgment that the pupil is working at the level of that test. As the test, at any level, can include questions from the whole of the programme of study, a teacher must give consideration to a pupil's attainment across the whole programme of study before entering him or her for the test. QCA will try to ensure that, as far as possible, items are developed which are accessible to younger pupils, without compromising validity.

The pilot for single level tests has been extended to see how tests perform as an accountability measure and QCA is now well-placed to take forward the development of single level tests, using this test model.

The first three test sessions of the pilot for single level tests have posed substantial challenges, but have also provided a firm evidence base for the confirmed test model. There will, however, continue to be technical issues arising in the context of the development of single level tests according to the test model that will need addressing and a substantial programme of work will be required to ensure that the tests can be demonstrated to meet the regulatory criteria of validity, reliability, comparability, minimising bias and manageability.

Appendix A – Equating design for the June and December single level test pre-test

Reading			SLT June 2008					NCT anchors		New anchors						SLT December 2008				
	s3	s4	s5	s6	s7	KS2 A	KS3 A	KS2 L3/4	KS2 L4/5	KS3 L4/5	KS3 L5/6	KS3 L6/7	KS2/3 L4/4	s3	s4	s5	s6	s7		
Year 5 3325	675	620	620					620												
	225	385		385				385												
	125	142	142						142											
	125	142		142					142											
	125	149		149						149										
	125	111			111					111										
	200	243	243	243																
	125	132	132												132					
	675	698							698						698					
	225	257							257							257				
	125	133								133					133					
	125	156								156						156				
	125	158									158					158				
	125	134									134						134			
200	237													237	237					
Year 6 3825	225	188	188					188												
	450	486		486				486												
	225	338			338			338												
	125	119	119						119											
	125	104		104					104											
	125	138		138						138										
	125	134			134					134										
	200	159	159	159																
	125	209		209	209															
	125	93	93												93					
	125	137		137												137				
	125	117			117												117			
	225	227							227						227					
	450	507							507							507				
	225	258							258								258			
	125	125								125					125					
125	125								125						125					
125	125									125						125				
125	125										125						125			
200	398													398	398					
125	415														415	415				
Year 7 1975	225	242		242				242												
	225	241			241			241												
	225	252			252				252											
	250	275		275									275							
	125	136		136											136					
	225	282						282							282					
	225	279						279								279				
	225	262							262							262				
250	293												293	293						
Year 8 2575	225	243			243			243												
	125	118			118					118										
	125	130			130								130							
	125	101				101							101							
	125	126				126								126						
	250	42					42							42						
	125	94		94	94															
	125	154			154	154														
	125	75				75	75													
	125	159			159												159			
	125	140				140												140		
	225	316						316									316			
	125	197									197						197			
	125	298										298					298			
	125	157										157						157		
125	163											163					163			
125	189												189	189						
125	266														266	266				
Year 9 1750	125	145			145							145								
	125	125			125								125							
	125	131				131							131							
	125	63				63							63							
	250	82					82						82							
	125	116			116	116														
	125	102				102	102													
	125	245				245												245		
	125	24										24					24			
	125	121											121				121			
	125	107											107					107		
125	133											133					133			
125	129															129	129			
Total per subject	13450	14477	1696	2899	2686	1253	301	5008	1073	796	824	484	1170	609	568	1918	3165	3164	1340	0

Key

- Red Analysis to link SLTs and NCTs (reading only using NCT anchors)
- Black Analysis to link SLTs and NCTs (using newly-constructed anchors)
- Blue Analysis to create vertical scale on SLTs
- Green Analysis to equate June 08 and December 08 SLTs
- Pink Analysis to link KS2 and KS3 standards

Data required to link June tests only



Additional data required to link June and December tests

Writing	Desired	Achieved	SLT June 2008					New KS2 anchor	New KS3 anchor	SLT December 2008				
			s3	s4	s5	s6	s7	all levels	all levels	s3	s4	s5	s6	s7
Year 5 1525	125	145	145					145						
	250	282		282				282						
	125	138			138			138						
	200	238	238	238										
	125	135	135							135				
	125	143						143		143				
	250	280						280			280			
	125	138						138				138		
	200	233								233	233			
Year 6 2025	125	144	144					144						
	250	293		293				293						
	125	147			147			147						
	200	226	226	226										
	125	147		147	147									
	125	142	142							142				
	125	145		145							145			
	125	139			139							139		
	125	138						138		138				
	250	282						282			282			
	125	134						134				134		
		200	229								229	229		
	125	155									155	155		
Year 7 625	500	280						280	280					
	125	148		148							148			
Year 8 1700	200	276			276					276				
	200	265				265				265				
	200	85					85			85				
	100	139		139	139									
	100	112			112	112								
	100	29				29	29							
	100	143			143							143		
	100	143				143							143	
	200	270							270			270		
	200	248							248				248	
	100	165									165	165		
	100	147									147	147		
Year 9 1400	200	116			116					116				
	200	88				88				88				
	200	42					42			42				
	100	67			67	67								
	100	53				53	53							
	100	83				83							83	
	200	277							277			277		
	200	274							274				274	
		100	81										81	81
Total per subject	7275	7584	1030	1618	1424	840	209	2544	2221	1020	1637	1649	976	0

Key	
Black	Analysis to link SLTs and NCTs
Blue	Analysis to create vertical scale on SLTs
Green	Analysis to equate June 08 and December 08 SLTs
Pink	Analysis to link KS2 and KS3 standards

	Data required to link June tests only
	Additional data required to link June and December tests

Mathematics	Desired	Achieved	SLT June 2008					New anchors					SLT December 2008					
			s3	s4	s5	s6	s7	KS2 L3/4	KS2 L4/5	KS3 L4/5	KS3 L5/6	KS3 L6/7	KS2/3 L4/4	s3	s4	s5	s6	s7
Year 5 1125	125	174	174					174										
	125	138		138				138										
	125	141		141					141									
	125	150			150				150									
	125	151	151											151				
	125	137						137						137				
	125	143						143							143			
	125	142							142						142			
	125	129							129							129		
Year 6 1375	125	161	161					161										
	125	142		142				142										
	125	135		135					135									
	125	161			161				161									
	125	133	133											133				
	125	144		144											144			
	125	142			142											142		
	125	138						138						138				
	125	134						134							134			
	125	144							144						144			
Year 7 625	250	297		297									297					
	125	105		105										105				
	250	273											273	273				
Year 8 1750	125	123			123					123								
	125	160			160						160							
	125	171				171					171							
	125	137				137						138						
	250	272					272					272						
	125	138			138										138			
	125	140				140										140		
	125	138							138							138		
	125	141								141						141		
	125	133									133						133	
	125	66										66					66	
	250	290											290					290
Year 9 1500	100	141			141					141								
	100	141			141						141							
	100	162				162					162							
	100	128				128						128						
	200	232					232					232						
	100	125				125										125		
	200	239					239										239	
	100	109							109							109		
	100	133									133					133		
	100	116										116					116	
	100	117											117				117	
	200	239											239					239
Total per subject	6375	7347	619	1102	1156	863	743	1167	1144	511	1157	1482	570	559	1085	1072	697	768

Key			
Black	Analysis to link SLTs and NCTs		Data required to link June tests only
Green	Analysis to equate June 08 and December 08 SLTs		Additional data required to link June and December tests
Pink	Analysis to link KS2 and KS3 standards		

References

Baird, J-A & Dhillon, D. (2006) Qualitative expert judgments on examination standards: valid but inexact. AQA: Guildford. Internal report.

Green, S, Pollitt, A, Johnson, M & Sutton, P. (2003) 'Comparability study of pupils' writing from different key stages.' Paper presented at the British Educational Research Association conference, University of Edinburgh: 11–13 September 2003.

Mason, K. (2003) *Cross key stage comparability study*. Mathematics Test Development Team, QCA, London

NFER/UCLES (1998). Cross key stage comparability study: English reading.

Wood, R. (1991) *Assessment and testing*, Cambridge University Press.