A Review of the Pilot of the Single Level Test Approach

April 2011

Ofqual/11/4837

Contents

Executive summary

The Qualifications and Curriculum Authority (QCA), now the Qualifications and Curriculum Development Agency (QCDA), was commissioned by the Department for Children, Schools and Families (DCSF), now the Department for Education (DfE), to develop the single level tests (SLTs) that constituted one component of the Making Good Progress (MGP) pilot. They were designed as a possible replacement for the end of Key Stage 2 tests and were planned to be implemented 'on a national basis at the earliest opportunity subject to positive evidence from the pilot and to endorsement of this approach from the regulator' (DCSF, *The Children's Plan*, 2007b, p. 67).

Although the SLT pilot scheme was cancelled three years later, after the new UK Government took office in May 2010, we had completed much of our work to review the new approach. We decided to summarise and publish our provisional findings in order to promote a better understanding of the pilot scheme and to document its conclusions for future reference.

This report provides a review of the SLT approach, based on existing evidence collected throughout the pilot scheme, in relation to the purposes set out for the tests and the regulatory requirements they would have to meet. We acknowledge that the pilot was ended before it reached a stage of making final proposals for an alternative approach to statutory assessment in Key Stage 2 and so we were not formally asked to give our view but we believe that, in keeping with our commitment to transparency, we should publish a summary of the work we had undertaken up to the ending of the pilot. The main findings from the review are summarised in this report.

The single level tests and their purposes

The MGP pilot was proposed by the DCSF in January 2007 and officially started in September 2007 (see DCSF, 2007a). The pilot contained five strands with assessment figuring in two major components:

- the assessment for learning (AfL) component aimed to improve ongoing assessment practice, drawing on evidence from day-to-day learning, and to improve the accuracy of assessment judgements that were reported by teachers for each pupil in the pilot schools each term
- the single level test (SLT) component aimed to develop tests that could be used to confirm the level of attainment of pupils as judged by their teachers; in this sense, the SLTs were designed for 'when ready' use.

For the SLT component, the DCSF commissioned QCA, now QCDA, to develop and deliver the pilot of the SLT strand of the MGP pilot, involving ten local authorities, initially with schools from both primary and secondary sectors (Key Stages 2 and 3).

The primary purposes of the SLTs were (DCSF/QCDA, 2010; QCDA, 2010g):

- to confirm that a pupil is working within a particular National Curriculum level as judged by their teacher by providing an independent measure of attainment
- to provide data from the tests (i.e. results) for inclusion in the achievement and attainment tables (AATs) and to allow the monitoring of national performance standards over time against the public service agreements (PSA) targets.

The constructs assessed by the SLTs were:

- proficiency in English reading (the English reading tests)
- proficiency in English writing (the English writing tests)
- proficiency in mathematics (the mathematics tests).

Proficiency is defined in terms of attainment of learning outcomes specified within the Key Stage 2 National Curriculum programmes of study (PoS) for English (reading and writing separately) and mathematics.

The intention was that these constructs for each subject were assessed by the SLTs at all levels. The only difference between the tests was the difference in the level of difficulty. Tests at all levels sampled content at the appropriate level from the Key Stage 2 PoS. It was also decided that the SLTs should carry forward existing Key Stage 2 National Curriculum Test (NCT) standards, defined by the National Curriculum level descriptions for mathematics, English reading and English writing, and defined by the performance of year 6 pupils who take these tests.

Findings from the review of the pilot of the single level test approach

Because assessment results are used to make a variety of decisions about the pupils being assessed, they must be accurate, valid and fit for purpose. We have set out common regulatory criteria with regard to all National Assessments we keep under review, comprising the following:

- Validity: The assessment should generate results that provide a valid measure of the required knowledge, skills and understanding as defined by the assessment objectives.
- Reliability: The assessment should generate results that provide a reliable measure of pupil performance.
- Comparability: The assessment should generate results that are comparable in standards over time and between test sessions.

- Minimising bias: The assessment should minimise bias, differentiating only on the basis of all pupils' ability to meet National Curriculum requirements.
- Manageability: The assessment should be manageable.

Evidence collected from the SLT pilot scheme has been reviewed against these criteria in relation to the primary purposes set for the SLT approach. In addition, this review has also considered the evidence for the potential impact of the SLT approach.

The following summarises the main findings from this review.

Meeting the regulatory criteria

Validity

Aspects of validity of the SLT approach, including face validity, content validity and concurrent validity, were demonstrated during the pilot. Although the SLT standards setting and maintenance process was rigorous and involved the use of both statistical information and professional judgement, its validity was not fully demonstrated due to insufficient information about the composition of the SLT anchor tests and evidence concerning the stability of item parameters of the anchor tests.

Reliability

The SLT pilot has demonstrated that the results from the SLTs have adequate levels of reliability for tests of this nature.

Comparability

Appropriate and valid procedures involving the use of both item response theory (IRT) and classical test theory (CTT) for test equating were employed to link the SLT standards to the NCT standards and to ensure the continuity of standards over time or between test sessions. The equating process was rigorous and the analysis was thorough. However, additional evidence on the stability of item parameters of items in the SLT anchor tests would be needed to fully demonstrate the comparability of standards over time.

Minimising bias:

The SLTs were administered following established procedures and items used in them went through a thorough validation process to ensure that they were suitable for the target population with construct irrelevance variances minimised, which created fairness to pupils. Procedures were also in place to ensure that the tests were accessible to pupils with special educational needs.

Manageability

Given the nature of the pilot and the limited use of SLTs in an accountability context, further evidence would be needed to demonstrate fully that the SLTs would be manageable if they were to be scaled up and rolled out nationally as high stakes tests. This would include both the human and financial resources and the operational manageability required to develop the tests and to operate the system.

Meeting the primary purposes

Using SLTs to confirm pupils' National Curriculum levels as judged by their teacher

The SLTs were intended to provide an independent measure of a pupil's achievement. This was used to confirm that the pupil was working within a particular National Curriculum level as judged by their teacher. It is not easy to interpret the extent or nature of agreement between SLT results and pupil entry decisions, which were based largely on teacher assessment judgements. Further evidence would be needed to fully demonstrate that the SLTs could be used effectively to independently confirm a level as judged by their teacher.

Monitoring national performance standards over time

Monitoring national performance standards over time accurately requires both teacher entry decisions and the SLT results to be accurate. Further evidence would be needed to fully demonstrate that the SLT approach could meet this purpose adequately.

Impacts of the SLT approach

The SLT pilot has produced evidence to support some of the intended positive impacts. These include less emphasis on teaching to the test, contribution to a broader and more balanced curriculum, improved tracking and monitoring of pupil progress, and reduced stress experienced by pupils. However, it is not clear to what extent those positive impacts would be sustainable if the SLTs were to be used as high stakes tests for accountability purposes.

Further evidence would be needed to fully demonstrate that the intended positive impacts would be achievable and could be maintained and potential negative impacts could be minimised if the tests were to be used as high stakes tests.

It should be noted that the conclusions presented above were drawn from the existing evidence from the SLT pilot scheme provided to us. If the pilot scheme had continued, further analysis by the test development agencies may well have been conducted and further evidence produced, which might have resulted in conclusions different from those outlined above.

1 Introduction

1.1 What are the single level tests?

The SLTs were externally marked tests in mathematics, English reading and English writing for pupils in Key Stage 2. Unlike the statutory end of Key Stage 2 NCTs, each SLT covered only one National Curriculum level and was marked on a pass or fail basis (QCA, 2009; QCDA, 2010e, f, g). End of Key Stage 2 NCTs on the other hand cover levels 3 to 5 inclusive and also award a compensatory level 2 for those pupils who come within 3 marks of achieving level 3. Levels are awarded based on the marks achieved in relation to a set of agreed level 'threshold' scores.

Unlike the end of Key Stage 2 NCTs, the SLTs had two test windows each year (scheduled for June and December) and could be taken by pupils in any year group in Key Stage 2 (years 3 to 6 inclusive). There were single level tests for levels 3, 4, 5 and 6 and entry decisions were made by teachers, principally on the basis of current teacher assessment judgements of pupils' performance.

Schools in the pilot were advised to only enter pupils for tests when they had been either (PwC, 2010a):

- teacher assessed as progressing to the next National Curriculum level since their last externally reported assessment (i.e. Key Stage 1 National Curriculum Assessments or NCAs); or
- operating within the level being tested, at any sub-level, or at sub-level 'a' at the level below for which they are entered.[1]

The primary purpose of the SLTs was to produce an independent measure of achievement to confirm that a pupil is working at a particular National Curriculum level as judged by the teacher (DfE/QCDA, 2010; QCDA, 2010e, g). Related to this was a second purpose of providing data from the tests (i.e. results) for inclusion in the AATs and to allow the monitoring of national performance standards over time against the PSA targets (DfE/QCDA, 2010).

The SLTs were developed and delivered by the QCA on behalf of the DCSF and were piloted initially with 355 primary and around 70 secondary schools across ten local authorities in England (QCDA, 2010e). After the June 2008 cycle, the SLTs were no longer piloted in secondary schools (QCDA, 2010a; PwC, 2010a). When the pilot was extended in 2009, schools were informed that mathematics results would be used for accountability purposes and were given the opportunity to withdraw from

[1] For the December 2007 test window, teachers were advised only to enter pupils who were securely operating at the level of the test (sub-level 'b' or above) as opposed to at the threshold.

the pilot, which reduced the number of schools to 226; by the end of the pilot 225 schools remained (QCDA, 2010e; PwC, 2010b).

For the first two cycles of the pilot, tests were also developed and available for pupils in Key Stage 3 but, from December 2008, the SLTs were only for use in Key Stage 2.

1.2 Rationale for the development of SLTs

In January 2007, DCSF launched a consultation document, *Making Good Progress: How can we help every pupil to make good progress at school?* Following on from this consultation, the MGP pilot began in September 2007, running until July 2009 (National Assessment Agency (NAA), 2008). The pilot contained five strands, one of which was the development of SLTs. It was subsequently decided to extend the SLT strand until June 2010.

The AfL strand of the MGP pilot focused on wider formative assessment activities and teacher assessment supported by the use of the Assessing Pupils' Progress (APP) assessment criteria (PwC, 2010a).

Teacher assessment was an integral part of the SLT approach since teachers' decisions to enter pupils for the tests were principally based on their current judgements about pupils' attainment in relation to national standards. The evaluation of the tests has to have some regard to issues of accuracy in teacher assessment and the difference in the constructs being assessed by the two different forms of assessment. Teacher assessment can take account of a wide range of evidence in a variety of contexts over time and applies a 'best fit' model to the level descriptions whereas tests, by definition, have to assess on the basis of pupils' responses to a selected sample of the curriculum on a single occasion.

1.3 Key changes during the pilot

After the DCSF decision to cease the use of statutory Key Stage 3 NCTs in October 2008, the remit of the SLT strand reduced to Key Stage 2 only from the December 2008 test cycle onwards (QCDA, 2010a; PwC, 2010a). This reduced the number of participating schools to 355 (QCDA, 2010e). At this point, pupils in all schools in the pilot took both the NCTs in year 6 and, if entered for them, the SLTs.

Based on recommendations from the Expert Group on Assessment, the DCSF decided in September 2009 that SLT mathematics results would replace those from NCTs in mathematics for accountability purposes for June 2010 and that pupils in the pilot schools would no longer take NCTs in mathematics in year 6 (QCDA, 2009). Schools in the pilot were required to opt into this change in order to continue as part of the pilot, leaving 225 schools in the pilot as of December 2009 (QCDA, 2010a). The DCSF decided that, for accountability purposes, year 6 pupils in these remaining

pilot schools would have reported their highest mathematics level achieved in any SLT from the December 2008 cycle onwards (QCDA, 2010b).

1.4 The purposes of SLTs

The primary purposes of the SLTs, which were defined in early 2008 and refined subsequently during the pilot, were (DfE/QCDA, 2010; QCDA, 2010g):

- to confirm that a pupil is working at a particular National Curriculum level as judged by their teacher by providing an independent measure of attainment
- to provide data from the tests (i.e. results) for inclusion in the AATs and to allow the monitoring of national performance standards over time against the PSA targets.

In addition to these primary purposes, a number of additional possible uses of the data were identified. These included to:

- play a role in helping teachers to set expectations and targets (including progression targets) for pupils
- provide information about the test levels achieved by a pupil (including the dates on which they were achieved) at the time of transfer between teachers and schools
- help schools set overall targets
- monitor schools' performance in comparison with other schools
- provide information to Ofsted for use in inspections
- provide data to calculate progression indicators
- provide measures of the effectiveness of national programmes to raise performance
- provide information on the performance of sub-groups of pupils, for example by gender, by ethnicity, those eligible for free school meals, and those with special educational needs.

Even though one could consider these as secondary uses of the data, they came with the following caution (DfE/QCDA, 2010):

> The person/organisation using the data for any of these specific purposes will need to consider whether evidence supporting the use of the tests for their primary purpose also supports their use in these instances.

1.5 Definition of constructs to be assessed by the SLTs and performance standards

The constructs assessed by the SLTs were:

- proficiency in reading (the English reading tests)
- proficiency in writing (the English writing tests)
- proficiency in mathematics (the mathematics tests).

Here proficiency is defined in terms of attainment of learning outcomes specified within the Key Stage 2 National Curriculum PoS for English (reading and writing separately) and mathematics and by their associated attainment targets and level descriptions (see Newton, 2009a).

The intention was that these constructs for each subject were assessed by the SLTs at all levels and the only difference between the tests was the difference in the level of difficulty. Tests at all levels sampled content at the appropriate level from the Key Stage 2 PoS. The assumption of a single construct across different levels is important as it allows for straightforward comparison of performances on tests at different levels (see later discussions) in the same subject. A clear definition of the construct to be assessed by an assessment is crucial for its subsequent development and evaluation.

In order to allow continuity of monitoring against public service targets, it was decided that the SLTs would inherit the standards from current National Curriculum Key Stage 2 tests as defined by the National Curriculum level descriptions for the attainment targets for mathematics, English reading and English writing respectively (QCDA, 2009, 2010a; QCA, 1999a, b).

1.6 Our role and activities in the pilot scheme

We were involved in the pilot from the outset as an observer, initially as the Regulations and Standards group of what was then QCA, prior to the creation of Ofqual. Our activities during the pilot were on an observer basis and included:

- observation of marker training (team leader and marker)
- observation of ‘live’ test administration
- observation of test development meetings
- observation of ‘live’ test construction
- observation of level confirmation meetings

- attendance at QCA/QCDA SLT programme board meetings
- attendance at the Assessment and Test Development Board
- attendance at meetings of the Test Delivery Group.

1.7 Purpose of this report

While QCDA had been commissioned to develop the SLTs that constitute one component of the assessment model of the MGP pilot, as a regulator we would have been requested by DCSF to endorse the approach to SLT had it been proposed for use as part of the national assessment arrangements (DCSF, 2007b). Although the SLT pilot scheme was cancelled three years after the new UK Government took office in May 2010, we had completed much of our work to review this new approach and decided to summarise and publish our provisional findings in order to promote a better understanding of the pilot scheme and to document its provisional conclusions for future reference. This report provides an independent review of the SLT approach, based on existing evidence collected throughout the pilot scheme, in relation to the purposes set out for the tests and the regulatory requirements they would have to meet. We acknowledge that the pilot was ended before it reached a stage of making final proposals for an alternative approach to statutory assessment in Key Stage 2 and so we were not formally asked to give our view but we believe that, in keeping with our commitment to transparency, we should publish a summary of the work we had undertaken up to the ending of the pilot.

2 Methodology

2.1 Scope of the review

As indicated earlier, assessment was at the heart of two major components of the MGP pilot:

- the AfL component, which aimed to improve ongoing assessment practice, drawing on evidence from day-to-day learning and to improve the accuracy of teacher assessment judgements which were reported by teachers for each pupil in the pilot schools each term
- the SLT component, which aimed to develop tests that could be used to confirm the level of attainment of a pupil as judged by their teacher; in this sense, the SLTs were designed for 'when ready' use.

Because entry for the SLTs was based principally on the current teacher assessment judgement of the pupil's performance and because the SLT results were intended to be used for performance monitoring purposes, the accuracy of the outcomes for individual students and for schools related to the accuracy of both the teacher entry decisions and the SLT results. Therefore, a review of the piloted approach should consider the assessment arrangements as a whole, including both the SLT and the reported teacher assessment judgements.

Because there was no specific evaluation of the accuracy of teacher assessment, this report has focused mainly on a review of the effectiveness of the SLTs themselves. However, since pupil entry for the SLTs was based largely on current teacher assessment judgement, some aspects of teacher assessment have also been considered. Further, since the pilot has gone through various changes, this review primarily focused on existing evidence collected from the test cycles from June 2009 onwards. The review does not examine the effectiveness of the SLT delivery system.

2.2 Review criteria and evidence gathering

Assessment results are used to make a variety of decisions about the pupils being assessed. We have set out common regulatory criteria in relation to the national assessment arrangements it regulates. These criteria comprise:

- Validity: The assessment should generate results that provide a valid measure of the required knowledge, skills and understanding as specified by the assessment objectives.
- Reliability: The assessment should generate results that provide a reliable measure of pupil performance.

- Comparability: The assessment should generate results that are comparable in standards over time and between test sessions.
- Minimising bias: The assessment should minimise bias, differentiating only on the basis of all pupils' ability to meet National Curriculum requirements.
- Manageability: The assessment should be manageable.

These criteria have been used to review the performance of the pilot SLT system in relation to the primary purposes set for the SLT approach. This review also considered early evidence of the potential impact of the SLT approach.

To gather the necessary evidence for judging how well the SLT pilot met the regulatory criteria, information from a range of sources was collected for analysis. These sources included:

- reports produced by the test development teams of QCDA (see Appendix A for a list of the reports received from QCDA)
- technical and research reports commissioned by QCDA and DfE (or DCSF) from test development agencies and other relevant organisations (see Appendix A for a list of the reports received from QCDA)
- other relevant reports in the public domain
- reports produced by us in relation to SLTs
- other published and unpublished relevant research reports/papers.

3 Technical quality of the SLT approach

The review of the technical quality of the pilot SLT approach considers whether the SLT component of the MGP pilot assessment approach was likely to produce results that were capable of supporting decision-making in relation to its intended purposes. This relates to the review of some aspects of validity of the SLT approach and would involve analysing evidence from the processes of test development, test administration, test scoring, score reliability, test equating, standards setting, comparability of standards over time or between test forms, attention to fairness to test takers and other aspects of the SLT approach (see AERA *et al.*, 1999).

3.1 The SLT development process

Effective test design and development is crucial for developing valid tests.

AERA *et al.* (1999) identifies four phases or steps involved in test development:

- Delineation of the purposes of the test and the scope of the construct or the extent of the domain to be assessed: This delineation involves the development of the test framework that extends the original statements of purposes and the construct into an elaboration that delineates aspects (e.g. knowledge, skills, content and etc.) of the construct or domain to be measured.

- Development and evaluation of test specifications: Based on the test framework, test specifications are developed and evaluated. Test specifications are the basis for the design of the test, which delineate the format of items and tasks, the response format and the scoring procedures including scoring rubrics (mark schemes). Test specifications may also consider factors such as the desired psychometric properties of items, time restrictions, target populations and procedures for test administration. They should also indicate how test scores will be interpreted (norm-referenced interpretation or criterion-referenced interpretation). Test specifications are used to guide the development of items and test construction.

- Development, field testing, evaluation, and selection of items and scoring guides and procedures: Test items and the associated mark schemes are developed based on the test specifications that identify the content domain and item format. The development and evaluation of test items and the scoring rubrics may involve the use of a range of interested and skilled personnel. Qualities of items are ensured through item review procedures and pilot testing.

- Assembly and evaluation of the test for operational use: The items selected for a test should meet the requirements of the test specifications. The construction of a test may consider factors like content quality and scope, the weighting of items and sub-domains and appropriateness of the items for the target

population. Any test should also be evaluated for other properties such as consistency of scoring procedures with the purposes of the test and differential item functioning (DIF).

Test construct framework

A clear statement of the purposes of the test to be developed and a clear definition of the construct to be measured by the test are important for developing valid tests. Although the SLT constructs have been defined as proficiency in mathematics, English reading and English writing in terms of the attainment of learning outcomes specified by the entire Key Stage 2 National Curriculum PoS, a test construct framework for each of the three subjects has yet to be developed to exemplify the construct to be assessed in the context of single level testing (QCDA, 2010e). The SLTs were used to confirm that pupils were working at a particular attainment levels as judged by their teachers. Decisions about pupil entry were based mainly on recent teacher assessments and teachers in the pilot were encouraged to use the assessment guidelines (criteria) developed for the APP framework as a basis for their termly judgements. Although both the SLTs and teacher assessment are based on the entire Key Stage 2 PoS, the two methods are distinctively different (see later discussions).

A test framework would ensure more consistent interpretation of the construct being assessed and the performance standards being applied between the SLTs and the teacher assessment. The test framework and its use in the subsequent test development process lay foundations for generating evidence to support the validity arguments proposed for the intended use of the results from the test.

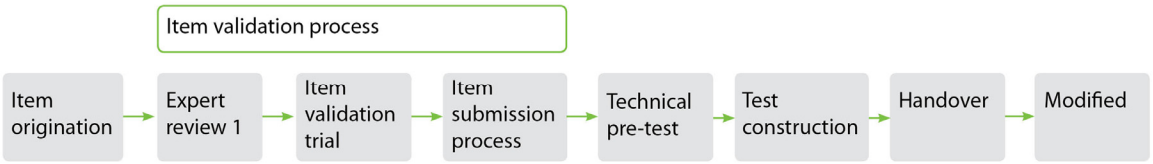
Test specifications

Although a test construct framework was not fully established, test specifications for the SLTs were developed, which drew on the Key Stage 2 National Curriculum PoS and the APP framework of level-related criteria, based on the national curriculum level descriptions, developed by QCA to support teacher assessment. The specifications were detailed in terms of sub-domain weightings, curriculum coverage and item types.

Item development and validation

Initially, a process similar to that used for developing National Curriculum tests was used to develop the SLTs. However, that was revised and Figure 1 depicts the revised new pilot SLT development process (QCDA, 2010e, f).

Figure 1 The SLT development process (adapted from QCDA, 2010g).



This new process was partially implemented for the first time in the June 2009 mathematics test. Items were created by experienced experts and a thorough validation process was undertaken to ensure their quality. The item validation process involved (QCDA, 2010e, f, g):

- First expert review: Items were evaluated by review panels that consisted of experts from a variety of relevant areas (curriculum, teachers, local authorities, inclusion, cultural issues and markers).
- Item validation trial: Items were trialled to ensure that they were functioning as required. This was to check that the items could be understood by pupils, the mark schemes were appropriate, the assessment focuses were appropriate and the items were accessible to pupils for the intended ages.
- Second expert review: The trialled items were further scrutinised by the review panels to ensure that the items met the acceptance criteria for technical pre-testing.

Technical pre-testing

Accepted items were pre-tested through a well-defined and rigorous technical process. The technical pre-test was designed to:

- evaluate the quality of the accepted items further and to produce item statistics that were used to effectively design the SLTs for live testing so that they are targeted at the specific levels of pupils with appropriate difficulty
- distinguish appropriately between pupils who were working at the levels being assessed and those who were below the levels.

Further, the technical pre-testing process also involved equating the newly constructed tests and the new SLT anchor tests directly so that the SLTs constructed for live testing were of known statistical properties and had provisional level thresholds set (see later discussions).

Construction of live SLTs

The item statistics (such as the IRT item difficulty and discrimination parameters) were used for selecting items from the technically pre-tested item pool to construct live SLTs. Because the item parameters were known, the properties of the constructed SLTs became known (assuming that the items were unidimensional and the assumptions of the IRT model that were used to analyse the pre-test data were met). In addition to statistical properties such as the average difficulty and maximum information near the level thresholds, other constraints such as content balancing were also considered. The known properties of the live SLTs before they were administered represent an important feature of the SLT test development process.

3.2 The SLT anchor tests

Test equating (under both CTT and IRT frameworks) plays an important role in maintaining standards over time for NCTs. Crucial to the process of test equating in NCTs is the use of anchor tests to link standards of performance over time and to account for pre-test effect (Donahue *et al.*, 2008; Bechger, 2008; Lin *et al.*, 2009; Maycock *et al.*, 2009; Pyle *et al.*, 2009; Whetton, 2009).

Pre-test effect refers to the existence of differences in the performance on a test between students taking the test under pre-test conditions and those taking the same test under live test conditions (taking into account differences in ability between the students). For the SLTs, IRT was the preferred equating tool for the mathematics and the English reading tests while CTT was used for equating the English writing tests. Tests that are used for equating should ideally be built to the similar content and statistical specifications (Kolen and Brennan, 2004, see also later discussions). It was also decided that the SLTs should carry forward existing Key Stage 2 NCT standards.

The construction of SLT anchor tests that are linked to the existing NCTs would seem to be an effective way to achieve this. For the level 6 SLTs, the standards were those of the then Key Stage 3 standards. QCDA commissioned two research reports investigating how level 6 should be conceptualised in the context of the Key Stage 2 PoS (see Pollitt and Ahmed, 2009; Warwick *et al.*, 2009). If continuing, further work would be needed to agree on the standards for SLTs at level 6 in the context of the Key Stage 2 PoS. The level 6 tests used in the pilot were not linked to the Key Stage 2 NCTs but were trialled informally in high-performing non-pilot schools (QCDA, 2010g). Were level 6 tests to be used as an option within the Key Stage 2 NCTs, there would be similar issues with setting the level 6 standards in each subject.

The SLT anchor tests

Unlike general qualifications, NCTs have been using secure anchor tests and pre-testing to maintain standards over time. It was also intended for SLTs to use anchor tests for maintaining standards over time. Five SLT anchor tests were created:

- mathematics: an SLT level 3/4 anchor test and level 4/5 anchor test composed of items from the 2007 live NCT
- English reading: an SLT level 3/4 anchor test and level 4/5 anchor test composed of items from the 2007 live NCT
- English writing: the anchor test used for linking all the levels was a modified version of the Key Stage 2 English writing anchor test for NCTs (the NCT anchor with the spelling part removed)
- the mathematics and English reading anchor tests were for 50 minutes and the English writing anchor test was for 65 minutes. For the mathematics and English reading anchor tests, the level 4 items were common items between the anchors. It would appear that the maximum available marks for the mathematics and English reading anchor tests were 40 and 27 respectively. The maximum mark for the English writing anchor test was 43 (Bechger *et al.*, 2008; Lin *et al*., 2009).

These anchor tests were also used to link the SLT standards to the existing NCT standards (see later discussions).

The anchor tests for the mathematics and English reading tests were constructed to link the SLT with an appropriate subset of NCT items in order to overcome issues associated with linking a SLT with a multi-level NCT or a multi-level National Curriculum anchor test (QCA, 2009).

IRT analysis was carried out on the live 2007 NCT data in order to place item difficulty and pupil ability onto the same scale. Level thresholds were then applied to classify items into different levels, in the same way as pupils can be classified into different levels. The anchor tests were examined by test developers to ensure they were appropriate in terms of content coverage and the items which were included.

Composition of the anchor tests

The newly created SLT mathematics and English reading anchor tests were composed of the 2007 live Key Stage 2 NCTs. A number of issues arose from this:

- Exposure of the SLT anchor items: Since the mathematics and English reading anchor items were derived from previously administered live NCTs, they had been exposed to schools in England, including those involved in pre-testing the SLT items. Since schools frequently use past tests as practice tests to prepare their students (see for example, PwC, 2010b), the statistical properties of the anchor items may not be stable. Instability of anchor item parameters can introduce systematic error in test results.

- It is not clear why the SLT mathematics and English reading anchor tests were not developed based on the existing Key Stage 2 NCT anchors given that the NCT anchors have been used for equating in pre-testing Key Stage 2 NCTs. The use of SLT anchors based on the existing NCT anchors would take into account pre-test effect automatically when equating and ensure continuity of Key Stage 2 NCT standards (see later discussions). The existing NCT anchor tests are used under secure conditions when pre-testing.

3.3 Technical pre-testing

Once items had been trialled and validated, they went through a pre-test process (see QCA, 2009; QCDA, 2010e, g).

Purposes of technical pre-test

The technical pre-test served the following main purposes, being designed to:

- establish accurate statistical properties of the items
- align the item difficulties with those of the SLT anchor tests (equating)
- construct SLTs using pre-tested items for live testing
- establish the provisional level thresholds for the SLTs (cut scores), i.e. to set and maintain standards which were equivalent to those set for the NCTs (carrying forward the standards set for Key Stage 2 NCTs).

Basic principles of test equating

In many situations such as the NCTs and SLTs, multiple forms or versions of the same test are used for reasons of security. Although different forms are administered to different groups of test takers at different times, it is constantly required that the results from all test takers are compared for the purpose of fairness. This is normally achieved through the process of test equating, which establishes mathematical relationships between the test scores from different forms so that they can be used interchangeably in terms of the level of achievement inferred from test scores regardless of which form of the test a test taker has taken.

Procedures involved in test equating generally include:

- developing test forms
- developing test data collection designs
- administering tests and collecting responses

- establishing the relationship between scores from different forms using statistical procedures to transform scores on one form to equivalent scores on a different form.

Kolen and Brennan (2004) (also see Lord, 1980; Petersen *et al.*, 1989; Yu and Popp, 2005) outlined some of the desirable properties of equating relationships. These include the following:

- Same specification: The two test forms to be equated are built to the same content and statistical specifications, so that they measure the same underlying trait or construct.
- Equity: For examinees of identical performance level on the underlying trait or true score, the conditional frequency distribution of scores on form Y (including observed score means and standard deviations), after transformation, must be the same as the conditional frequency distribution of scores on form X.
- Group invariance: The equating relationship must be the same regardless of the group of examinees from which it was derived.
- Symmetry: The equating transformation must be symmetric or invertible, that is, the mapping of scores from form X to form Y must be the same as the mapping of scores from form Y to form X. This requires that the function used to transform scores on form X to the form Y scale be the inverse of the function used to transform scores on form Y to the form X scale.

Bèguin (2000), Kolen and Brennan (2004) and Livingston (2004) discussed some of the most widely used equating designs and their advantages and disadvantages in CTT and IRT equating. These include:

- the single group design
- the counterbalanced design
- the equivalent groups design
- the internal anchor design
- the external anchor design
- the pre-equating non-equivalent groups design and other designs.

These designs can also be conceptualised as two broad designs: common item design and common person design.

Procedures for test equating under both the CTT and the IRT frameworks have been developed. IRT equating is more flexible in terms of equating designs for linking test forms, but it is complex, both conceptually and procedurally. While CTT uses directly raw score transformation between test forms, IRT equating is based on an abstraction of the underlying person trait being measured by the test. The use of common items or common persons for equating using IRT involves similar procedures, which are to place item parameters from different test forms on the same common scale.

The SLT technical pre-testing process

The SLT technical pre-test process included a complex equating design for the mathematics and English reading tests, taking into consideration item order effect, sample size and ability, and other factors. This pre-testing process involved the following procedures:

- Item blocks were formed of the accepted new items that were at different difficulty levels. Each block was similar to a SLT.
- These item blocks and the SLT anchor tests (also referred to as blocks) were used to form test booklets (each test booklet may contain a single block or two blocks).
- These booklets were administered to year 6 pupils selected from schools outside the MGP pilot. Pupils were assigned to different ability groups who then took the appropriate booklets. Each block of items was taken in one test session, and the order of the blocks being taken met a counterbalance equating design.
- The booklets were then marked and data collected for analysis together with the data from the live 2007 NCTs from which the anchor tests were derived in the case of mathematics and English reading.
- In the case of the English writing test, paired SLTs or SLTs and the anchor tests were administered to year 6 students and the data was collected for analysis for the December 2009 and June 2010 test sessions (for previous test sessions, students from different year groups were used for equating).

QCDA also commissioned a research project looking at alternative methods, particularly the embedding of items in live testing, to pre-test items. A report was produced that discussed the possibilities and issues associated with live test embedding (Cito, 2009b; also see Wheadon *et al.*, 2009).

Samples used for technical pre-testing

In order to carry forward the NCT standards, samples of pupils used for pre-testing were year 6 students for whom the full Key Stage 2 PoS had been covered. Reasonable sample sizes were achieved for equating analyses.

Analysis of technical pre-test data – equating

Both CTT and IRT procedures were used to analyse the technical pre-test data.

For the mathematics and English reading SLTs, an extension of the Rasch model (referred to by the test equating agency as the 'one-parameter logistic model' or OPLM) was used for data analysis. In OPLM discrimination, parameters are imputed as known integer constants, having been estimated in an early stage of analysis and then converted to integer constants with a specified geometric mean. This model is therefore argued to represent a compromise between the attractive mathematical properties of the Rasch model (in particular, the ability to use conditional maximum likelihood estimation, which makes no assumption about the underlying ability distribution of the samples) and the flexibility of the two parameter logistic model (namely, the ability to take the differing discriminatory power of items into account). A concurrent calibration approach was used, involving analysing the pre-test data collected from all booklets and the NCT live test data in order to align item difficulty and person ability parameters onto a common scale.

For the recent test cycles, both CTT and IRT analyses on the SLT pre-tests were thorough, including (see Bechger, 2008, 2009a,b, 2010; Bechger and Bèguin, 2009; Donahue *et al.*, 2008; Bechger and Maris, 2009; Lin *et al.*, 2009; Maycock *et al.*, 2009):

- basic item and test statistics
- item order effect
- pre-test effects
- DIF and item bias
- local dependence of items
- item model fit statistics
- random equating error.

Since the anchor tests were derived from the 2007 live NCTs, a concurrent approach to calibrating items from all the test booklets for mathematics and English reading resulted in automatic equating between the SLTs and the 2007 live NCTs. Pre-test effect was taken into account by removing items in the anchor tests that showed DIF

between 2007 live testing and SLT pre-testing from the linking items and the application of a concurrent calibration approach. The exclusion of anchor items showing DIF from the linking items also took into account differential pre-test effect associated with some anchor items and represented an improvement over the equating procedures currently used for NCTs.

For English writing SLTs, a number of CTT equating procedures, including equipercentile (with and without post-smoothing) and linear equating, were used to link the SLTs with the anchor test under pre-test conditions, based on performances of the sample pupils on the SLTs and the SLT anchor test and the performances of a sample on the full NCT anchor test used for equating the Key Stage 2 English writing test (see Lin *et al.*, 2009).

For the level 6 SLTs, tests were developed as whole test papers by the test development agency. They were only informally trialled in schools before being used for live testing.

Construction of live tests and setting provisional level thresholds

Test construction meetings, involving the test development agencies and QCDA, were held to construct final SLTs for live testing by selecting pre-tested items to meet test specifications and maximum measurement information (for mathematics and English reading SLTs from levels 3 to 5), considering the following requirements (see Cito, 2009a; Bechger, 2009a, b, 2010; QCA, 2009; QCDA, 2010e):

- approximately 40 per cent of the marks was made available for attainment at the level below the level of the test and the remaining 60 per cent for attainment at the level of the test (i.e. approximately 40 per cent of the mathematics items are in common between two adjacent SLTs)
- maximum marks available on the test
- balance of content coverage
- balance of different types of items
- cut scores were about 50 per cent of the maximum available marks
- maximum test information at cut scores
- all SLTs took 60 minutes to complete (for the June 2010 test session).

Because the items were calibrated onto the same common scale of the anchor tests and the ability level thresholds set for the NCTs were used for the SLTs, once a SLT was constructed, the provisional level cut scores were automatically set. The attempt

to set provisional cut scores near the middle of the targeted ability range for the SLTs was appropriate.

In the case of the English writing test, the pre-tested tests were used for live testing, and the provisional cut scores were those determined by the pre-testing results (Donahue *et al.*, 2008; Lin *et al.*, 2009; Maycock *et al.*, 2009).

Carrying forward the standards of the end of Key Stage 2 tests

It was decided at the outset of the pilot that SLTs should carry forward standards for the Key Stage 2 NCTs. The use of past live test items to construct the new SLT anchor tests and the inclusion of the 2007 live NCTs with known level thresholds in the analysis was intended to achieve this for mathematics and English reading (see Bechger, 2008; Bechger *et al.*, 2008). For English writing, this was achieved through linking the SLT anchor test with the existing full Key Stage 2 NCT anchor test with known level thresholds.

Although conceptually both NCTs and SLTs measure the same constructs and sample contents from the whole of the Key Stage 2 PoS, there is a difference between them operationally and this has implications for the operational definition of the National Curriculum levels and the equivalence of standards for NCTs and SLTs. The inference of a student's attainment level from NCTs is based on their overall performance on all items from different levels in the test (a compensatory assessment model) and the items are analysed holistically. In contrast, SLTs were designed to use items from two adjacent National Curriculum levels to make a judgement about whether or not a student has achieved a particular level, essentially implying a weak mastery assessment model. Further, some of the skills such as those required for the mental test in mathematics and spelling test in English writing, assessed in NCTs, were not assessed in SLTs or assessed in a different way.

Further evidence needed

The equating designs for the mathematics and English reading SLTs were quite complex, taking into consideration factors such as item order effect, sample size and ability. Further evidence in the following areas would be useful:

- Characteristics of the samples taking the 2007 live NCTs: No detailed descriptions of the samples who took the live 2007 NCTs were provided.
- The mental mathematics test: For the mathematics tests, it appeared that the analysis of pre-test data involved the use of the full 2007 live NCT data that contains mental mathematics test data. The SLTs in mathematics do not include mental tests. Although the anchor items derived from the live NCTs do not include the mental test items, it is not clear what the impact would be on the SLT level thresholds. For the English writing SLTs, the anchor test is a revised version of the full Key Stage 2 NCT anchor test (with the spelling section

removed), and the equating involved establishing the level thresholds on the SLT anchor test by equating the SLT anchor with the full Key Stage 2 NCT anchor. It would seem that a similar procedure would be appropriate for the 2007 live Key Stage 2 NCT (at least to assess the impact of removing the mental test).

- Test of IRT model assumptions: For some of the test sessions, results from more detailed analysis, including local dependence and model fit statistics, were reported. The use of unidimensional IRT models in test data analysis also requires that the items in the tests measure a unidimensional latent variable. Given that the tests covered a wide range of ability and various levels of knowledge and skills, it is important to examine how well the tests met this assumption, because the properties such as sample invariant of item parameters depend on the test data meeting the model assumptions sufficiently. Although the item and person fit statistics to a certain degree reflect unidimensionality, they are normally not very sensitive to departure from unidimensionality (Smith, 2002; Zoanetti *et al.*, 2009).

- Comparability – provisional level ability thresholds and cut scores: It appeared that for each test cycle for mathematics and English reading, the same analysis procedures, involving the use of 2007 live Key Stage 2 National Curriculum tests, were adopted. This raises the question as to whether the item difficulty scales established using the NCTs for different cycles were the same, given that each cycle had different pupil samples and different SLT pre-tests. This is because it is very unlikely that the different tests were measuring exactly the same latent variables. More evidence would be needed to demonstrate that the scales were the same.

- Stability of anchor tests and anchor item parameters: This is also closely related to comparability. There was no information available on the stability of the anchor tests in terms of the items used for equating in different test series (items showing DIF between SLT pre-testing and 2007 NCT live testing were not used as link items). It is not clear what the implications for comparability of standards would be if different anchor items were used for equating in different test cycles. If the anchor tests were stable, it would be useful to examine the consistency of the anchor tests in terms of expected cut scores (or parameter estimates) over the different test sessions if the anchor items were calibrated using both the 2007 live NCT data and the SLT pre-test data, and variation of the expected cut scores could be an indication of inconsistent or non-comparable standards. The stability of anchor item parameters (both individually and as a group) between test sessions is important for equating to be accurate and for continuity of standards to be maintained through test equating (see discussions above). Further, if the NCT sample size was substantially larger than those for the SLT pre-tests and the anchor tests for

each test cycle and a concurrent calibration approach was used, the anchor item parameter estimates would be determined predominantly by the NCT responses. A comparison of the performances of the pre-test samples on the anchor items between test cycles would be useful.

- Equating error: Analysis on random equating error was conducted for some of the test cycles. Discussion about possible systematic error and error sources would be useful.

3.4 Live testing and analysis

Test administration

The live SLTs were administered in accordance with the standard NCT procedures. Appropriate quality assurance procedures were applied.

Tests were administered securely by the schools under the direction of their own staff. How this was organised varies by school, subject and according to the number of pupils. Generally, from our observations, undertaken alongside QCDA colleagues, tests were taken in the classroom with pupils sitting all levels together. In some cycles, due to very small number of pupils, level 6 tests were securely administered by QCDA staff so that the test could be reused in a later cycle.

Marking and marking quality monitoring

Unlike NCTs, where marking is paper based, the SLTs were all marked onscreen using onscreen marking software. The candidates' scripts were electronically scanned and then uploaded into the software that markers then used to mark the scripts onscreen. Other than in very rare cases, the marker never saw the actual paper copy of the candidate script.

The marking process for the SLTs was rigorous and thorough. Marking quality was assured through (QCDA, 2010e, f):

- Marker quality: Detailed specifications of criteria were developed for the selection of the three types of markers (level leaders, team leaders and markers).

- Marker training: The markers were trained for consistent use of mark schemes. Appropriate marker training materials were developed for the training.

- Marking quality assurance: Two quality assurance mechanisms were use to ensure marking quality: standardisation and seeding. The standardisation process involved comparing marks given to a set of questions by markers with the definitive marks assigned during the standardisation and seeding meetings to examine whether the pre-defined tolerance level was met. Seeding was used

to monitor live test marking quality and involved comparing marks given to seed questions by markers with definitive marks.

- Alternative approach to marking and standard setting for SLT in English writing: An alternative approach to marking and standard setting and maintenance for SLT in English writing using paired comparative judgement was explored. The initial results appeared to be encouraging (see Pollitt *et al.*, 2009).

Live test analysis

The purposes of analysing live test data were to establish (Bechger and Donahue, 2009a, b):

- the performance of the tests under live test conditions
- the final, definitive pass scores.

Comparison of performance of SLT items between pre-testing and live testing
Comparison analyses of the performance of the SLT items and the tests between pre-testing and live testing were conducted (Bechger and Donahue, 2009a, b). For mathematics and English reading tests, live SLT data were analysed with SLT pre-test data to identify items showing DIF between pre-testing and live testing. SLT items showing DIF were not used for establishing the common scale between the SLTs and the 2007 live NCTs. This again took into account differential pre-test effect associated with some SLT items between pre-testing and live testing.

Entry patterns
For all the subjects, only a small proportion of the pupils entered for the tests at all levels were from year 3. Most pupils who took the level 3 tests were from year 4 and year 5, a large proportion of those who took the level 4 test were from year 5, and those who took the level 5 test were mainly from year 6. These patterns to a certain degree reflect the levelled progression in teaching and learning (QCDA 2010e, g).

Pass rate at each level by year groups
For the June 2009, and December 2009 and June 2010 test cycles, the pass rates showed substantial variability between subjects and between year groups at different levels (QCDA, 2010e, g). This variation in the pass rates may have reflected variability in teacher assessment judgement (unreliability), leading to incorrect test entry decisions or inappropriate SLT standards or a combination of both.

Relationship between test scores and teacher assessment sub-levels
Test scores were positively related to reported teacher assessment sub-levels. This was expected, given that the pupil entry for the SLTs was largely based on their current teacher assessment level. The pass rates at each level generally increased with increasing teacher assessment sub-levels. This again was expected (QCDA, 2010e, f, g).

Performance on the SLTs by levels, gender and year group.
Detailed analysis was conducted on SLT performance at each level by different year groups and gender (QCA, 2009; QCDA, 2010e, f, g).

Marking reliability
For the December 2009 test session, a detailed marking reliability study was conducted on all levels of the English reading and English writing tests, with sample size about 200 at each level (QCDA, 2010e). The study investigated the consistency in assigning marks to questions/ranking pupils and in classifying pupils into pass or fail categories. For the English reading test, the consistency in classifying pupils into pass or fail categories varied from 89 per cent for level 5 to 99 per cent for level 4. For English writing, the consistency ranged from 72.6 per cent for level 5 and 98 for level 3. Overall, results from the study indicated that the mark reliability of single level English reading and writing tests was comparable with that of assessments of a similar nature. Approaches were suggested to improve marker reliability.

Test reliability and reliability indices for SLTs
Reliability is broadly defined as the consistency of results on a given measure from repeated measurements under equivalent conditions. Reliability is an important indicator of the quality of an assessment under the CTT framework. Depending on how the measurement procedure is repeated, there are different forms of reliability, including test–retest reliability (the same test is taken at different times by the same test takers), alternative or parallel form reliability (different forms of the same test are taken by the same test takers) and marker reliability (the same test scripts are marked by different markers).

The reliability of test scores is normally measured by the reliability coefficient representing correlation between two sets of scores. However, many high stakes tests are administered only once and the response data is frequently used to produce a measure of reliability referred to as the internal consistency reliability. The internal consistency reliability represented by Cronbach's alpha essentially involves splitting a test into two halves and correlating the scores on the two halves and reflects how consistent different set of items in a test produce similar or consistent scores. There are many factors that can affect the reliability of test scores and the reporting of any reliability indices should indicate the magnitude of the indices and the sources of errors that they account for.

QCDA commissioned two research projects investigating how reliability should be conceptualised in the context of SLTs (see Johnson and Johnson, 2008; Bramley, 2009).

Reliability coefficients are related to raw scores or scaled scores and are useful for comparing tests or measurement procedures. However, when test results are reported using raw scores or scaled score or performance categories such as the teacher assessment levels, the reliability coefficients are difficult to interpret. In the

case of raw scores or scaled scores, the standard error of measurement (SEM) that is related to the reliability of the test and the standard deviation of the raw scores or scaled scores is frequently used. In the case of reporting performance categories, classification indices are used. For SLTs, the results are reported using National Curriculum levels and classification indices are useful. There are two types of classification indices:

- classification accuracy, which refers to the consistency of the classification by the observed test scores and the true scores of the same test
- classification consistency, which refers to the consistency of classification by two sets of observed scores (of the same test or of two different tests).

Table 1 shows the internal reliabilities of the SLTs in mathematics and English reading for the June 2009 and December 2010 test cycles (QCDA, 2010e).

Table 1 Internal reliability measures for two test sessions.

Subject	Level	Cronbach's alpha	
		June 2009	December 2009
Mathematics	3	0.88	0.86
	4	0.82	0.83
	5	0.78	0.80
English reading	3	0.77	0.80
	4	0.70	0.72
	5	0.64	0.63

Values of Cronbach's alpha are similar for the two test sessions. As expected, these values are generally lower than those reported for the corresponding NCTs because the SLTs are substantially shorter than the NCTs and reliability generally increases with test length (see Newton, 2009c; Hutchison and Benton, 2009; Maughan *et al.*, 2009). Furthermore, the variation of ability in pupils is less for the SLTs than that for NCTs, while reliability can increase with increasing variation of ability in the test takers.

The classification accuracy indices for the SLTs in mathematics and English reading were estimated under the IRT framework and the values are shown in Table 2 (QCDA, 2010e). These values seem relatively high in comparison with those for NCTs. However, it is realised that classification accuracy is a function of the mark distribution, level threshold position and the standard error of measurement. Furthermore, for NCTs, there are five performance categories (National Curriculum levels) and four level boundaries.

Table 2 Classification accuracy indices for two test sessions.

Subject	Level	Classification accuracy (%)		
		June 2009	December 2009	June 2010
Mathematics	3		92	96
	4		93	90
	5		90	88
English reading	3	95	93	91
	4	95	95	91
	5	93	86	89

Analysis on meeting IRT model assumptions and item fit statistics
For the December 2007, and June 2009 and December 2009 test cycles, additional analysis of the IRT model fit to the live test data was conducted (Schagen *et al*. 2008; Bechger and Donahue, 2009a, b; QCDA, 2010e). Analysis of unidimensionality and local independence assumptions of the IRT model used was also conducted. For the mathematics and English reading SLTs in the December 2009 cycle, the unidimensionality of the OPLM model held reasonably well, but there were a few items for which local independence assumption was violated. There were also a few items that were not discriminative.

Construct irrelevant variance analysis of the SLTs in English reading
This construct irrelevant variance (CIV) analysis was conducted to identify the main sources of construct irrelevant difficulty and easiness that were present in the SLTs in English reading for the June 2009 and December 2009 test cycles (QCDA, 2010f). The analysis investigated different sources of CIV for the different types of questions, question–mark scheme consistency and mark scheme differentiation for constructed response questions. Results from this analysis indicated that most of the items appeared to perform as intended. Findings from the analysis could be used to improve future test development.

Other basic test and item statistics
For the December 2009 test session, some basic item statistics, including item facilities and item-total score correlation, were also provided. These statistics generally showed that the items were functioning reasonably well (see Schagen et al., 2008; QCA, 2009; Bechger and Donahue, 2009a, b; QCDA, 2010e, f). It appeared that analysis of DIF between pre-testing and live testing was also carried out.

3.5 Level thresholds setting and standards confirmation

For SLTs, maintenance of standards over time or between test sessions was through maintaining the continuity of the underlying latent traits being measured by the tests using test equating to establish level thresholds on SLTs, which involved aligning item parameters of SLT items to the scale of the anchor tests. Ideally, once the latent trait level thresholds for individual levels on the anchor tests have been established, they need to remain stable during their lifetime. Although the ability level thresholds may stay unchanged in order to maintain standards, the corresponding cut scores on individual SLTs can vary from test session to test session as a result of variation in item difficulties and lengths between tests composed of pre-tested items.

For the SLTs in mathematics and English reading:

- the provisional cut scores were automatically established once a live test had been constructed using pre-tested items
- live SLT data was used to establish the final definitive cut scores.

Level confirmation

This is conducted by all four marking level leaders (levels 3 to 6) from the same test (i.e. all reading level leaders for reading, etc.). For levels 3 to 5, the four level leaders looked at scripts at the statistically identified threshold and at one mark either side and decided whether they could or could not accept the threshold. It is important to note that for levels 3 to 5, the level confirmation exercise was to confirm the cut score and not to set it (QCDA 2010e, g).

As there were insufficient numbers of pupils at level 6 in the pilot, a statistical threshold could not be identified by equating and statistical analysis. Therefore for level 6 tests the level leaders conducted a cut score identification exercise. This involved the level leaders reviewing scripts individually from a variety of scripts across the mark range to identify those that they felt provided sufficient evidence that the pupil was working at level 6 and those that did not. For reading and writing, due to the small number of entries, it was possible for level leaders to have access to all scripts within the range. As mathematics had a larger number of entries, the level leaders looked at a range of scripts centering on the intended cut score, 50 per cent of the available marks (QCDA 2010e, g).

Standards confirmation

The SLT standards confirmation process involved a panel of members from the test development team, psychometrics team, marking and other functional delivery teams to consider the quality of the evidence collected from technical pre-testing and live testing, and outcomes from the level confirmation exercise to recommend the final cut scores (QCA, 2009; QCDA, 2010e, g). These meetings were observed by us. The

primary evidence for levels 3 to 5 was the statistical evidence that took precedence over the evidence from the qualitative level confirmation exercise. For level 6, no statistical evidence was available so only the qualitative cut score identification evidence can be considered (QCDA, 2010e).

Standards confirmation is where the final level thresholds are agreed, and recommended to QCDA's accountable officer for sign-off (QCDA, 2010g).

3.6 Use of SLT results to confirm pupils' attainment as judged by their teachers

The primary purpose of the SLTs was to provide an independent measure to confirm that a pupil was working at a National Curriculum level as judged by their teacher.

Use of SLTs to measure student attainment levels

As indicated earlier, the constructs assessed by the SLTs are proficiency in mathematics, English reading and English writing. The Key Stage 2 PoS embody a wide range of knowledge and skills constituting the National Curriculum. The use of results from these tests to measure students' levels of attainment in the relevant subjects requires that the tasks contained in them are representative samples from the domain of similar tasks (the assessed domain), which in turn is representative of the domain of all potential tasks (the target domain) representing the whole Key Stage 2 PoS. Evidence of the representativeness of the tasks in the tests and the representativeness of the assessed domain has to be demonstrated for the results to be valid (see Crooks *et al.*, 1996; Crisp, 2009).

Teacher assessment judgements in the MGP pilot

Teacher assessment was a focus of the other major component of the MGP pilot, which dealt with assessment. Under the heading AfL, it aimed to improve the practice and use of ongoing classroom assessment to support planning and provision. It also encouraged the use of APP materials, which had been previously developed to support more systematic assessment of knowledge, skills and understanding in each National Curriculum attainment target in English and mathematics. To assist teachers to make professional and consistent judgements, pilot schools were provided with assessment guidelines in the form of a grid of assessment criteria that detailed the performance at different levels in a number of assessment focuses. For each attainment target, teachers were encouraged to make a level judgement in each assessment focus and could then reach an overall level judgement for the target. In the MGP pilot schools, teachers were required to submit a teacher assessment each term to provide a judgement of the National Curriculum level at which each pupil was performing in relation to mathematics, English reading and English writing. Judgements were submitted by sub-level where 'a' was high, 'b' was secure and 'c' was low.

The reliability of teacher assessment

Teachers made entry decisions for the SLTs mainly on the basis of their judgements about the particular National Curriculum levels at which pupils were operating. This means that the reliability and accuracy of both teacher assessment and SLTs would need to be considered as part of a full evaluation. There has been only limited research into the reliability of the results from teacher assessment due to operational difficulties and complexities in conducting that kind of research (see Wilmut, 2005; Stanley *et al*., 2009). Although the APP approach is intended for teachers to use consistent criteria in assessing their pupils and to produce consistent results, there has been no published evidence about the extent to which this has been achieved. Stanley *et al*. (2009) suggested various procedures for conducting teacher assessment reliability studies. Based on a comparison between the National Curriculum levels assigned to pupils who took the Key Stage 2 mathematics pre-test and their final National Curriculum levels (sample size 9,856, the two sets of data were collected a few weeks apart), Stanley *et al*. (2009) reported that 81 per cent of the pupils were assigned the same attainment levels by their teachers. The correlation between the two assessments was estimated to be 0.81. Hutchison and Benton reported an agreement rate of 66 per cent between the levels assigned based on an English pre-test and the levels assigned by teachers. Stanley *et al*. (2009) also reported reliability measures for various teacher assessments from different countries, and they showed substantial variability, with values as low as 0.39 being observed.

Interpretation of disagreement between SLTs and teacher assessment judgements

Results from the pilot test sessions indicated that the pass rate, or the level of agreement between the entry decisions, based largely on results from teacher assessment and SLT results, varied substantially between test sessions and between levels for the three subjects. For NCTs, teachers assign pupils to different tiers consisting of several levels but for SLTs, pupils were assigned to just a single level and this would require that teachers make a more accurate estimate of pupils' achievement. In the case of a lower pass rate, it was suggested that teachers may have made inappropriate entry decisions. However, the disagreement between teacher judgements and SLT results needs to be interpreted with caution:

- Although both the SLTs and the APP criteria on which teachers base their assessment of pupils are assumed to measure the construct of the whole Key Stage 2 PoS, the approaches are distinctively different. While teachers base their judgements on observations and classroom interactions over time, SLTs employ written tasks at a specific point. As indicated earlier, the domain assessed by the SLTs only represents part of the target domain. The same will be true for teacher assessment although one would expect it to draw on a wider domain. Differences between the two assessed domains can therefore exist in

terms of differences in the emphasis of the knowledge, skills and understanding perceived to reflect the learning requirements of the curriculum by the two methods. The discrepancy or disagreement in outcomes may therefore be legitimate and to some extent is related to the limitations in the accuracy of results from any assessment procedure.

- In addition to the possible differences in the constructs measured by the two assessed domains, both teacher assessment and SLT results have errors (see previous discussions) that could produce misalignment between the two sets of results.

Areas for further evidence

As indicated earlier, evaluation of a system of tests for which pupils are entered mainly on the basis of current teacher judgement requires consideration of the extent to which the two assessments measure the same construct and are equally reliable. Additional evidence would be required in the following areas in order to explore the relationship between the two approaches.

Reliability of teacher judgements
It is important to investigate the reliability of any assessment since reliability is a prerequisite of validity. A comprehensive study of the reliability of teacher judgements would be required in order to investigate the likely impact of unreliability in teacher assessment results (and consequently on tests entry decisions) on the agreement rate between SLTs and teacher assessment.

Justification of the use of SLTs to confirm teacher assessment judgement
As discussed earlier, results from both teacher assessment and the SLTs will have errors and these have implications for the relationship between the two assessments. Previous research has suggested that for the Key Stage 2 NCTs in science, the consistency between two tests (equivalent) was generally below 80 per cent (Maughan *et al.*, 2009). For English reading, work by Hutchison and Benton (2009) suggested a consistency of about 70 per cent between the levels based on the NCT test and the levels assigned by teachers. Given unreliability in results from both SLTs and teacher assessment, care would be needed in interpreting very high or very low pass rates among the pupils entered for the tests.

3.7 Results reporting

Once the pass marks were decided at the Standards Confirmation meeting based on statistical (equating) evidence (not available for level 6) and professional judgement (level confirmation; cut score identification for level 6 only) and signed off by QCDA's accountable officer, results were communicated to schools. Results were made available via the Key to Success website (www.keytosuccess.dcsf.gov.uk) for both schools and local authorities. It was the school's responsibility to communicate

results to pupils and their parents/guardians. Evidence suggests that this was not widely done (PwC, 2010a).

Similarly, all schools were able to access item level feedback, although evidence from the first year of the pilot suggested that it was not widely shared (PwC, 2010a). Whether this improved during the later cycles has not been reported.

All pupils who were entered for a SLT were eligible for a certificate of participation, although anecdotal evidence indicates that some schools chose not to pass these onto the pupils.

3.8 Performance standards monitoring over time

One of the main objectives of the SLTs was that the results would be able to be used to monitor national performance standards over time. Because the SLT results would have been used to make inferences about the attainment levels of pupils for the entire Key Stage 2 PoS, their successful use to monitor performance standards over time would require the following:

- High reliability: The tasks assessed are representative of the tasks from the assessment domain so that scores from the assessment can be generalised to the universe scores for the assessed domain.
- High validity: The assessed domain is a sufficient representation of the target domain so that the appropriate knowledge and skills are assessed to ensure that the universe scores for the assessed domain can be generalised to the universe scores for the target domain.
- High comparability: Although initially the performance standards may be set arbitrarily and subjectively, they should remain constant afterwards.

Reliability

The single level tests have demonstrated adequate levels of reliability.

Validity

The item creation process for the SLTs was rigorous and the items went through a thorough validation process. Although test specifications were developed and the tests were supposed to draw tasks from the full range of the Key Stage 2 PoS, an analysis of the tasks assessed by the tests in relation to the full Key Stage 2 PoS would be useful for establishing the representativeness of the assessed domain. As indicated by Crooks *et al.* (1996), (also see Messick, 1989, 1995; AERA *et al.*, 1999), construct under-representation remains an important threat to validity. Some aspects of validity of the SLT approach have been demonstrated (see later discussions).

Comparability

As with National Curriculum tests, test equating has been used to maintain the comparability of standards over time. For the SLTs in mathematics and English reading, a unidimensional IRT model (the OPLM) was used for test equating, while the CTT approaches were used for equating English writing tests.

Appropriate and valid procedures for test equating were employed to link the SLT standards to the NCT standards and to ensure the continuity of standards over time or between test sessions. The equating process was rigorous and the analysis was thorough. However, additional evidence on the stability of item parameters of items in the SLT anchor tests would be needed to fully demonstrate the comparability of the SLT standards over time.

3.9 Access issues, reviews and script archiving

The procedures for the development of the single level tests were similar to those adopted for National Curriculum tests. The issues associated with assessment arrangements and appeals were generally addressed properly. The online marking of the tests made archiving of scripts very easy and efficient.

Access arrangements

Single level access arrangements fall into three categories (QCDA 2010c):

- those delegated to schools to use at their own discretion
- those that require a notification form to be returned with the pupils' test script
- those that must be applied for and approved by QCDA prior to the test being undertaken.

Even in light of these access arrangements, the PricewaterhouseCoopers (PwC) evaluation of the MGP pilot found that pupils with special educational needs were more likely to be eligible but not entered for a SLT than pupils without (PwC, 2010a).

Analysis by ethnicity of the proportion of pupils eligible but not entered for a SLT found that those from Black African and from any other Black background were most likely to be eligible but not entered (PwC, 2010a). The pupils least likely to be eligible and not entered were Indian and Bangladeshi pupils (PwC, 2010a).

Reviews

Reviews were first introduced for the December 2009 cycle for mathematics. This was expanded to all subjects for the June 2010 cycle. Reviews were carried out by senior markers and involved reviewing the marking of the whole test script (QCDA,

2010d). As with NCTs, schools only paid for unsuccessful reviews. Statistics on review outcomes were included in national statistics.

Script archiving

One of the major advantages of the SLTs was that, as marking occurred onscreen, all scripts were scanned which allowed for electronic archiving.

4 The impacts of the SLT approach

Any national educational assessment system will have direct or indirect impacts on the main stakeholders, the national educational system and the wider society in general. In addition to the positive impacts intended for an assessment, there can also be unintended negative impacts. The positive consequences should outweigh the negative consequences in order for the assessment system to be beneficial. Unintended negative consequences present serious threats to the validity of the intended use of the results from the assessment system.

4.1 Intended positive impacts of SLTs

The following positive impacts were presumed by DCSF to follow from an assessment approach based upon confirmatory, 'when ready' SLTs (see DfES, 2007a, b; DCSF 2007b; DCSF 2008; Hansard 2008 and HoC, 2008; also see Newton, 2009b):

- will motivate and accelerate progress more than present tests (i.e. the tests – which are intended to celebrate achievement – will establish more, and therefore less distant, goals)
- will be more useful to parents and pupils than present tests
- will be no more burdensome than present tests (i.e. more frequent but shorter)
- students will be more likely to experience success than on present tests (i.e. less likely to face questions that are beyond their capability)
- will not encourage 'teaching to the test'
- will make the test experience feel less 'high stakes' for pupils
- will contribute to better teaching and learning
- will reduce the need for other tests to monitor progress within the Key Stage or for diagnostic purposes.

In addition to these SLT-specific presumed positive impacts, the combination of teacher assessment and SLTs were also presumed by DCSF to have the following positive impacts:

- make formative assessment and summative assessment work together more effectively
- strengthen the relationship between ongoing teacher assessment and formal testing

- permit more effective use of group teaching
- smooth transitions into and between schools
- facilitate a more personalised learning experience.

DCSF (now DfE) commissioned PwC to evaluate the SLT pilot which focused on the areas of test entry, test preparation and revision, impact on schools, and impact on pupils, parents and carers (PwC, 2010b). Findings from the evaluation provided evidence to support some of the positive impacts, particularly in the areas of less emphasis on teaching to the test, contribution to a broader and more balanced curriculum, improved tracking and monitoring of pupil progress, and reduced amount of stress experienced by pupils. However, it should be realised that the SLT pilot was of low stakes (with the exception of the June 2010 mathematics tests for year 6 pupil). There was no information available on the representativeness of the findings as the number of schools that participated in both phase 1 and phase 2 of the survey is not known. It is also worth noting that for both phases of the surveys the participating schools accounted for 50 per cent and 48 per cent of the schools in pilot, but about one third of the schools that started the pilot. It would be interesting to know why these schools withdrew from the pilot and if their experiences were the same. Further, some of the findings were from focus group discussions that may not be generalisable. Further work would be needed to produce evidence for other intended positive consequences.

4.2 Investigating potential unintended impacts

It was proposed that, if adopted nationally, the SLTs results would be used as a measure for school accountability, making the tests high stakes (the pilot was low stakes except for mathematics in year 6 in the final stage of the pilot). As with any high stakes tests, there can be both intended and potential unintended consequences (Madaus *et al.*, 2009). Some of the unintended consequences can be negative. Most of the pilot was conducted in a low stakes context except for mathematics in year 6 in the final stage of the pilot and so evidence of these potential unintended negative impacts is limited.

The following areas would need to be investigated for potential negative impacts if the SLTs were to be used as high stakes tests.

Teaching to the test and narrowing of the curriculum

The evaluation of the SLT pilot by PwC suggested that the inclusion of year 6 mathematics results in the AATs resulted in an increase in the amount of time spent by schools in test preparation and revision for the subject (PwC, 2010b). It was also found, along with the earlier PwC report, that there was no evidence of teaching to the test in the pilot schools (PwC, 2010a, b). However, it has to be recognised that

during the pilot, SLTs were not used for accountability purposes except for mathematics in year 6 in the later stages of the pilot. Would using the SLTs more widely for accountability purposes encourage teaching to the tests and hence narrow the curriculum?

Test anxiety for pupils

Would SLTs put pupils under additional pressure to prepare for the tests and produce increased test anxiety for some pupils?

Pupil motivation and self-efficacy

While passing a SLT may give some pupils a sense of achievement and therefore motivate them to learn and make further progress, would the use of levelled tests create more frequently 'levelled' pupils and negatively affect low achieving pupils that are 'resitting' tests?

The role of teacher assessment

The use of the SLT results as the only accountability measures could mean that teacher judgements would only be used to make SLT entry decisions. What impact would this have on the role and importance of teacher assessment?

Impact on teaching and the curriculum

Because the Key Stage 2 PoS spans a period of four years and the SLTs sample knowledge and skills from the entire Key Stage 2 curriculum, a pupil could pass a test prior to the completion of the entire curriculum. This is not possible with the current NCT arrangements. What impact does this difference have on the curriculum and teaching and learning?

Consequences of inappropriate test entry and unreliability of SLT results

What would be the consequences of inappropriate test entry and re-entry or the potential unreliability of SLT results? How would it affect teacher behaviour and pupils' learning and progression?

Burden on schools, teachers and pupils

It is not clear how the more frequent testing of the SLTs would not place a larger burden on schools, teachers and pupils than the existing National Curriculum tests. The shorter time required to complete SLTs would probably not be able to explain this, given that pupils in the same class might need to be registered and prepared for different tests at different levels. What would be the burden of the SLTs? How would that compare with the existing NCTs?

Financial costs of implementing the SLT approach

The higher frequency of testing and the number of SLTs required each year would be demanding in terms of item creation, recruitment of representative samples for pre-testing, analysis of technical pre-testing data and construction of live tests, administration of tests, marking and administration. What would be the financial costs of implementing the SLT approach? Would the SLT system represent value for money?

5 Overall review summary

The previous sections looked at the technical quality of the piloted SLT approach and the intended and potential unintended impacts it could have had on the main stakeholders and the national education system. This section provides a brief summary of the degree to which the SLT approach, as piloted, met the regulatory criteria in relation to its primary purposes.

5.1 Meeting the regulatory criteria

Validity

Validity refers to the extent to which theory and evidence support the interpretation of assessment results for the intended uses of the assessment and is multi-faceted. The focus of this review has been placed on the following aspects of validity:

Face validity
Face validity refers to the extent to which a test and the results are perceived to be accurate and appropriate by the main stakeholders. The existence of rigorous item validation and test development processes ensured that SLTs had a high level of face validity.

Content validity
Content validity refers to the degree to which a test represents adequately the domain of content for the construct to be assessed. In the case of SLT, the content domain represents the full Key Stage 2 PoS. The use of relatively detailed test specifications and the involvement of expert judgements in the item creation and test development processes for the SLTs ensured content validity.

Standards setting
The standards setting process for SLTs was rigorous and involved the use of both statistical evidence and professional judgements, although the emphasis was placed on statistical evidence. However, information on the stability of item parameters for the SLT anchor tests over time was not available and some technical issues have been identified regarding the composition of the SLT anchor tests that would need to be addressed for the process to be fully valid (see later discussions on comparability).

Criterion-related validity
The relatively high pass rate for the SLTs suggested good agreement between SLT results and entry decisions (strongly influenced by teacher assessment judgements), indicating evidence of concurrent validity.

To summarise, most aspects of validity of the SLT approach have been demonstrated. However, further evidence would be needed to fully demonstrate the validity of SLT results for their intended uses.

Reliability

Reliability refers to the consistency of test scores from repeated measurements under equivalent conditions. Reliability coefficients and decision accuracy and consistency indices have been appropriately used in the context of the SLTs. The reliability of assessment results is affected by a range of factors, from test construction to standards setting.

Consistency in test administration
The live SLTs were administered in accordance with the standard NCT procedures. Appropriate quality assurance procedures were applied.

Consistency in marking and marking reliability
For SLT English reading and English writing tests, particularly writing, inconsistency in marking will present the major factor introducing unreliability in test scores. Detailed mark schemes for SLTs were developed that were used to guide test marking. The marking of the SLTs followed that for NCTs, using stringent procedures for training markers and monitoring the quality of marking. Marker reliability analysis indicated that the marker reliability measures for SLTs were comparable with those for NCTs for the same subjects.

Test item-related reliability
High stakes tests generally involve the use of equivalent or alternative forms of the same test, and each form consists of items or tasks sampled from the assessed domain. Longer tests generally have higher level of reliability than short tests because the potential overlap of similar questions between two long test forms will likely to be larger than between two short test forms. Because the SLTs are significantly shorter than the NCTs, the internal consistency reliabilities as represented by Cronbach's alpha for SLT in mathematics and English reading were generally lower than those for the NCTs.

Classification consistency and accuracy
For the multi-level NCTs, results are reported using performance categories represented by National Curriculum levels. When results are reported using performance categories, classification indices (both classification accuracy and classification consistency) are more appropriate and easy to interpret than the reliability coefficients for raw scores. For the pilot SLTs, pupils either passed a test or failed it, and classification accuracy for the Mathematics and English reading tests and classification consistency for English reading tests have been analysed to be relatively high. This would be expected since SLTs only report two performance categories, substantially less than those for NCTs.

Consistency in setting standards
For the SLTs, the established procedures were used for standards setting and maintenance (see discussions below).

To summarise, the pilot SLTs produced adequate levels of reliability for assessments of this nature.

Comparability

The consistency and accuracy of standards setting is the key to maintaining the comparability of standards over time or between test sessions. Although expert judgements were considered when setting pass scores for a SLT, the primary evidence was from equating through the technical pre-testing process and analysis of live test data. The SLTs were designed to carry forward standards from the Key Stage 2 NCTs and the use of items from the 2007 live NCTs to construct anchor tests for linking the SLTs and the NCTs was intended to achieve this. Both IRT and CTT equating procedures were used appropriately for analysing test data to link the SLT standards to the NCT standards and to ensure the continuity of standards over time or between test sessions.

Composition of anchor tests and their representativeness
For mathematics and English reading, the SLT anchor tests were constructed using the 2007 Key Stage 2 live NCTs, which would suggest that the items have been exposed to schools in England, including those involved in pre-testing SLT items. Detailed analysis on the rationale for the selection of the items to include in the SLT anchors and the representativeness of the items in terms of the knowledge and skills to be measured by the assessed domain was not available. For the English writing tests, the SLT anchor test was a revised version of the full Key Stage 2 NCT anchor test.

Stability of anchor item parameters
The accuracy of standards setting and maintenance for SLTs relied on equating the SLT pre-tests (and the live test) with the SLT anchor tests. This in turn required that items in both the pre-tests and the anchor tests met the various model assumptions and that the item parameters of the anchor items were stable between test sessions. For the mathematics and English reading tests, information on the stability of anchor item parameters (both individually and as a group) between test cycles was not available (the anchor items were calibrated using both the 2007 live NCT data and the SLT pre-test data for each test cycle). It is also not clear whether the anchor tests were stable in terms of the same items used as linking items for each individual test session.

To summarise, appropriate and valid procedures were employed to link the SLT standards to the NCT standards and to ensure the continuity of standards over time or between test sessions. The equating process was rigorous and the analysis was

thorough. However, additional evidence on the stability of item parameters of items in the SLT anchor tests would be needed to fully demonstrate the comparability of standards over time.

Minimising bias

SLTs were administered following standard procedures and items used in the test went through a thorough validation process to ensure that they were suitable for the target population with construct irrelevant variances minimised, which created fairness to pupils. Procedures were also in place to ensure that the tests were accessible to pupils with special educational needs.

Manageability

It is not clear that the SLTs would be manageable if they were to be scaled up and rolled out nationally as a high stakes test system in terms of both human and financial resources required to operate such a system (see previous discussions). Further, the SLTs were pre-tested using schools outside the pilot scheme involving the use of large samples of pupils; it could be difficult to conduct the existing item pre-testing process if they were taken by all schools.

5.2 Meeting the primary purposes

Using SLTs to confirm pupils' national curriculum levels as judged by their teacher

The precise relationship between the outcomes of SLTs and the teacher judgements about the pupils' overall attainment which informed their entry decisions was not fully explored in the pilot.

- Although both the SLTs and teacher assessment might be assumed to measure the same construct and apply the same performance standards when judging pupils, their operational definitions are different. While teachers make their overall judgements based on classroom interactions and other behaviours of the pupils from a broad perspective and over time, SLTs are, like any test, based on performance on a set of items/tasks sampled from their assessed domain on a single occasion. The difference in assessed domains and sampling of tasks between SLTs and teacher assessment may well make the operationally assessed construct different in the two approaches.

- It is possible that, as the pilot progressed, teachers began to make entry decisions for pupils for a SLT based on criteria perceived to be defined for a level by the SLTs. They may also have been influenced by the results of pupils who had been entered in previous test windows. In other words, the overall teacher judgement for a pupil might be different from the judgement made for test entry decisions. If this is the case, then the SLTs might be said to have

been used to confirm the pupil's level based on teacher entry decision. This obviously raises the question of how the correct performance standards should be operationally defined.

- Unreliability exists in both SLT results and teacher assessment judgements, and this raises questions about the relationship between SLT results and teacher assessment judgements. If it was true that both SLTs and teacher assessment measure a similar construct, then SLT results may be used to validate teacher assessment judgements or vice versa.

- Given a degree of unreliability in both teacher assessment judgements and SLT results, there are difficulties in interpreting the SLT pass rate or the disagreement between SLT results and entry decisions based largely on teacher assessment. If a large disagreement exists, which of the approaches is inaccurate, the SLT results or the teacher assessment judgements?

It was not clear from the pilot exactly how SLT outcomes would have related to teacher assessment judgements and how differences in the results from the two assessment approaches would be interpreted. It is worthwhile noting that externally-set tests and tasks are used at Key Stage 1 to inform statutory teacher assessment judgments but are not reported.

Monitoring national performance standards over time

Had they been implemented, the use of SLT results to monitor national performance standards over time in the piloted subjects would have depended on:

- both the SLT results and entry decisions, informed by teacher assessment judgements, being accurate

- the comparability of SLT standards being maintained.

From the discussions presented above, it is clear that, for the SLT pilot, further evidence would be needed to fully demonstrate that this objective could be achieved.

It is noted that the above conclusions were drawn from the existing evidence from the SLT pilot scheme provided to us. If the pilot scheme had continued, further analysis by the test development agencies may well have been conducted and further evidence may have been produced that could have resulted in different conclusions from those outlined above.

References

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (1999) *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Assessment Reform Group (2002). Testing, motivation and learning. Available at: www.assessment-reform-group.org/TML%20BOOKLET%20complete.pdf.

Au, W. (2007) High-stakes testing and curricular control: A qualitative metasynthesis. *Educational researcher*, **36**, 258–267.

Bechger, T. and Bèguin, A. (2009) *Equating the December single level tests: pilot 2008*. Cito.

Bechger, T. and Donahue, B. (2009a) *Report on the live testing of the July 2009 single level tests*. Cito and QCDA.

Bechger, T. and Donahue, B. (2009b) *Report on the live testing of the December 2009 single level tests: Final version*. Cito and QCDA.

Bechger, T. and Maris, G. (2009) *Stability of IRT-based true-score equating: Preliminary report*. Cito.

Bechger, T. (2008). *Equating single level tests: Results from the June 2008 data*. Cito.

Bechger, T. (2009a) *About the analysis of the technical pre-test data for the June 2009 single level tests in mathematics and reading*. Cito

Bechger, T. (2009b) *The analysis of the technical pre-test data for the December 2009 single level tests in mathematics and reading*. Cito.

Bechger, T. (2010) *Report on the technical pretest, and the test construction of the June 2010 single level tests in reading and mathematics*. Cito.

Bechger, T., Bèguin, A. and Verschoor, A. (2008) *Preliminary report on the selection of anchors for the single level test equating*. Cito.

Bèguin, A. (2000) Robustness of equating high-stakes tests. PhD thesis, University of Twente, the Netherlands.

Bramley, T. (2009) *Conceptualisation the reliability of single level tests.* A report to the NAA*. Cambridge Assessment*.

Cito (2009a) *Selecting the items for the June 2009 single level tests in mathematics: Preliminary report*. Cito.

Cito (2009b) *Alternative methods of obtaining item statistics for single level tests: embedding items*. Cito.

Crisp, V. (2009) Does assessing project work enhance the validity of qualifications? The case of GCSE coursework. *Educate*, **9**: 16–26.

Crooks, T., Kane, M. and Cohen, A. (1996) Threats to the valid use of assessments. *Assessment in education: Principles, policy and practice* **3**: 265–286.

DCSF (2007a) DCSF News: Making Good Progress: 10 local authorities selected for groundbreaking pilot. Available at: http://tna.europarchive.org/20070402211936/http://www.dfes.gov.uk/pns/DisplayPN.cgi?pn_id=2007_0046

DCSF (2007b) *The Children's Plan*. Nottingham: DSCF Publications. Available at: www.education.gov.uk/publications/standard/publication Detail/Page1/CM%207280

DCSF (2008). *The assessment for learning strategy*. Nottingham: DSCF Publications. Available at: publications.teachernet.gov.uk/eOrderingDownload/DCSF-00341-2008.pdf

DfE/QCDA (2010) *SLT purposes document*. Unpublished document.

DfES (2007a). *Making Good Progress: How can we help every pupil to make good progress at school? Consultation*. Nottingham: DfES Publications. Available at: www.education.gov.uk/publications/eOrderingDownload/DCSF-Making%20Good%20Progress.pdf

DfES (2007b). *Making Good Progress consultation: Government response*. DCSF Consultation Results website. Available at: www.dfes.gov.uk/consultations/downloadableDocs/Making%20Good%20Progress%20consultation%20-%20government%20response.doc

Donahue, B., Maycock, L. and Harries, L. (2008) *Evaluation of the June 2008 single level tests: Item analysis and equating*. A report submitted to the National Assessment Agency. NFER.

Hansard (2008). Commons Hansard Debates text for 9 October 2008, Volume No. 480 Part No. 142. Column: 148WH.

Herman, J. and Baker, L. (2009) Assessment policy: make the sense of Babel. In G. Sykes, B. Schneider and D. Plank (Eds) *Handbook of education policy research*. American Education Research Association. Routledge, New York, USA.

Hill, K. and Wigfield, A. (1984) Test anxiety: a major educational problem and what can be done about it. *Elementary school journal*, **85**, 105–126.

HoC CSF Committee (2008, July). *Testing and assessment: Government and Ofsted responses to the Committee's Third Report of Session 2007–08*. London: The Stationery Office Limited.

Hutchison, D. and Benton, T. (2009) *Parallel universes and parallel measures: Estimating the reliability of test results.* Ofqual, Coventry, UK. Available at: www.ofqual.gov.uk/files/2010-02-01-parallel-universes-and-parallel-measures.pdf.

Johnson, S. and Johnson, R. (2008) *Conceptualisation, establishing and monitoring of the reliability of single-level tests*. Assessment Europe Ltd.

Kolen, M. and Brennan M. (2004) *Test equating, scaling, and linking: Methods and practices*. Springer, Berlin.

Kruger; L., Wandle; C. and Struzziero' J. (2007) Coping with the stress of high stakes testing. *Journal of applied school psychology*, **23**, 109–128.

Lee, W. (2008). *Classification consistency and accuracy for complex assessments using item response theory* (No. 27). CASMA Research Report. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Levinson, C. (2000). Student assessment in eight counties. *Educational leadership*, **57(5)**, 58–61.

Lin, Y., Morrison, Jo. and Rutt, S. (2009) *Equating of the June 2009 writing single level tests: A draft report submitted to the Qualifications and Curriculum Authority*. NFER.

Livingston, S. (2004) *Equating test scores (without IRT)*. Educational Testing Service, New Jersey, USA.

Lord, F. (1980) *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum, New Jersey, USA.

Madaus, G., Russell, M. and Higgins, J. (2009) *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Information Age Publishing Inc., Charlotte, USA.

Maughan, S., Styles, B., Lin, Y. and Kirkup, C. (2009) *Partial estimates of reliability*. Ofqual, Coventry, UK. Available online at: www.ofqual.gov.uk/files/2009-11-partial-estimates-of-reliability-report.pdf

Maycock, L., Donahue, B. and Benton, T. (2009) *Equating of the December 2008 writing single level tests.* A final report submitted to the Qualifications and Curriculum Authority. NFER.

Messick, S. (1989) Validity. In *Educational measurement* (ed. by R. Linn), 13–103. Macmillan, New York, USA.

Messick, S. (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, **50**: 741–749.

NAA (2008) *What are single level tests?* Available at: http://webarchive.nationalarchives.gov.uk/+/www.naa.org.uk/naa_17894.aspx

Newton, P. (2009a). SLT constructs and standards. Unpublished Ofqual internal report.

Newton, P. (2009b) A framework for endorsing the single level test approach. Unpublished Ofqual internal report.

Newton, P. (2009c) The reliability of results from National Curriculum testing in England. *Educational research,* **51**, 181–212.

Petersen, N., Kolen, M. and Hoover, H. (1989) Scaling, norming, and equating. In *Educational measurement* (ed. by R. Linn), 221–262. Macmillan, New York, USA.

Pollitt, A., Derrieck, K. and Lynch, D. (2010) *Single level tests of KS2 writing: Paired comparisons research report*. TAG developments.

Pollitt, P. and Ahmed, A. (2009) *Conceptualisation of level 6 performance in the context of the key stage 3 programme of study*. Cambridge Exam Research.

Popham, W. (2001) Teaching to the test. *Educational leadership*, **58**, 16–20.

PwC (2010a) *Evaluation of the Making Good Progress pilot*. PricewaterhouseCoopers. ISBN 9781847756114. DfE report Research Report DCSF-RR065. Available at: www.education.gov.uk/publications/eOrderingDownload/DCSF-RR065.pdf

PwC (2010b) Evaluation of the single level test (SLT) pilot: Final report. DfE Research Report DFE-RR039. Available at: www.education.gov.uk/publications/eOrderingDownload/DFE-RR039.pdf.

Pyle, K., Jones, E., Willams, C. and Morrison, J. (2009) Investigation of the factors affecting the pre-test effect in national curriculum science assessment in England. *Educational research*, **51**: 269–282.

QCA (1999a) *English: The National Curriculum for England, key stages 1–4*. London, Department for Education and Employment and QCA.

QCA (1999b) *mathematics: The National Curriculum for England, key stages 1–4*. London, Department for Education and Employment and QCA.

QCA (2009) *Single level tests: Report of the first three test sessions: December 2007, June 2008 and December 2008*. QCDA.

QCDA (2009) *Single level tests pilot schools*. Available at: www.qcda.gov.uk/news/960.aspx

QCDA (2010a) Schools in the single level tests pilot. Available at: www.qcda.gov.uk/resources/assets/Pilot_schools_for_website.pdf

QCDA (2010b) Bulletin 2–March 2010: June 2010 test entries window. Available at: www.qcda.gov.uk/6487.aspx

QCDA (2010c) *Single level tests access arrangements guide for the June 2010 test cycle*. Available at: www.qcda.gov.uk/resources/assets/June_access_arrangements_guide.pdf

QCDA (2010d) Single level test Bulletin 8 July 2010. Available at: https://qca.custhelp.com/cgi-bin/qca.cfg/php/enduser/doc_serve.php?2=SLT-Bulletin8,

QCDA (2010e) *Single level tests. Report of the June 2009 and December 2009 test cycles*. QCDA.

QCDA (2010f) *Technical evaluation of single level tests in reading and writing: June and December 2009*. QCDA.

QCDA (2010g) *June 2010 single level tests: Report on development and outcomes*. QCDA.

Sacks, P. (2000) Predictable losers in testing schemes. *School administrator*, **57(11)**, 6–9.

Schagen, I., Maycock, L. and Benton, T. (2008) *The analysis of live single level test data*. NFER.

Select Committee Response to Government Response (2008) Available at: www.publications.parliament.uk/pa/cm200708/cmhansrd/cm081009/halltext/81009h0006.htm

Select Committee Testing and Assessment Inquiry (2008) Available at: www.publications.parliament.uk/pa/cm200708/cmselect/cmchilsch/cmchilsch.htm

Smith, E. (2002) Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of results. *Journal of applied measurement*, **3**, 205–231.

Stanley, G., MacCann, R., Gardner, J., Renolds, L. and Wild, I. (2009) *Review of teacher assessment: evidence of what works best and issues for development*. Report to QCA. Available at: www.education.ox.ac.uk/assessment/uploaded/2009_03-Review_of_teacher_assessment-QCA.pdf

Verstarlen, H. and Bechger, T. (2008) *Classification accuracy of educational tests*. Cito.

Volante, L (2004) Teaching to the test: What every educator and policy-maker should know. *Canadian journal of educational administration and policy,* **35**, September 25.

Warwick, I., Dickenson, M. and Worthen, L. (2009) *Report on conceptualisation of level 6 performance in the context of the key stage 2 programme of study*. London Gifted and Talented.

Wheadon, C., Whitehouse, C., Spalding, V., Tremain, K. and Charman, M. (2009) *Principles and practice of on-demand testing*. Ofqual, Coventry (accessed on 05 February 2010 at: Available at: www.ofqual.gov.uk/files/2009-01-principles-practice-on-demand-testing.pdf.

Whetton, C. (2009) A brief history of a testing time: national curriculum assessment in England 1989-2008. *Educational research*, **51**: 137–159.

Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education,* **3-13, 29 (3)**, 17–21.

Wilmut, J. (2005) *Experiences of summative teacher assessment in the UK*. A review conducted for the QCA. London: QCA

Yu, C. and Popp, S. (2005). Test equating by common items and common subjects: Concepts and applications. *Practical assessment research and evaluation*, **10(4)**. Available at: pareonline.net/getvn.asp?v=10andn=4.

Zoanetti, N., Griffin, P., Beaves, M. and Wallace, E. (2009) Rasch scaling procedures for informing development of a valid fetal surveillance education program multiple-choice assessment. *BMC medical education* 9:20. Available at: www.biomedcentral.com/1472-6920/9/20.

Appendix A SLT documents received from QCDA

Technical and equating reports

Bechger, T. and Bèguin, A. (2009). *Equating the December single level tests: pilot 2008*. Cito.

Bechger, T., Bèguin, A. and Verschoor, A (2008) *Preliminary report on the selection of anchors for the single-level test equating*. Cito.

Bechger, T. and Donahue, B. (2009a) *Report on the live testing of the July 2009 single level tests*. Cito and QCDA.

Bechger, T. and Donahue, B. (2009b) *Report on the live testing of the December 2009 single level tests: Final version*. Cito and QCDA.

Bechger, T. and Maris, G. (2009) *Stability of IRT-based true-score equating: Preliminary report*. Cito.

Bechger, T. (2008). *Equating single level tests: Results from the June 2008 data*. Cito.

Bechger, T. (2009a) *About the analysis of the technical pre-test data for the June 2009 single level tests in mathematics and reading*. Cito

Bechger, T. (2009b) *The analysis of the technical pre-test data for the December 2009 single level tests in mathematics and reading*. Cito.

Bechger, T. (2010) *Report on the technical pretest, and the test construction of the June 2010 single level tests in reading and mathematics*. Cito.

Cito (2009a) *Selecting the items for the June 2009 single level tests in mathematics: Preliminary report*. Cito.

Cito (2009b) *Alternative methods of obtaining item statistics for single level tests: embedding items*. Cito.

Donahue, B., Maycock, L. and Harries, L. (2008) *Evaluation of the June 2008 single level tests: Item analysis and equating*. A report submitted to the NAA. NFER

Lin, Y., Morrison, Jo. and Rutt, S. (2009*) Equating of the June 2009 writing single level tests*. A draft report submitted to the QCA. NFER.

Maycock, L., Donahue, B. and Benton, T. (2009) Equating of the December 2008 writing single level tests. A final report submitted to the QCA. NFER.

Patrick, H. (2009) *Single level tests: Level setting for December 2008*. Report to the Single Level Test Technical Advisory Group (TAG).

QCA (2009) *Single level tests: Report of the first three test sessions: December 2007, June 2008 and December 2008*.

QCDA (2010e) *Single level tests: Report of the June and December 2009 test cycles*.

QCDA (2010f) *Technical evaluation of single level tests in reading and writing: June and December 2009*.

QCDA (2010g) *June 2010 single level tests: Report on development and outcomes*.

Schagen, I., Maycock, L. and Benton, T. (2008) *The analysis of live single level test data*. NFER.

SLT Technical Advisory Group (TAG) (2008) *Single level tests: Interim technical report: December 2007 test session.*

SLT Technical Advisory Group (TAG) (2009) *Commentary and recommendations following the June 2008 single level test pilot*.

SLT standards confirmation documents

QCDA (2008) December 2007 —- English reading meeting notes
QCDA (2008) December 2007 – English writing meeting notes
QCDA (2008) December 2007 – mathematics meeting notes
QCDA (2009) June 2008 – Meeting notes
QCDA (2009) June 2009 – Standards confirmation presentation
QCDA (2009) June 2009 – Standards confirmation booklet
QCDA (2009) June 2009– Standards confirmation level setting report
QCDA (2009) December 2009 – Standards confirmation minutes
QCDA (2009) December 2009 – Standards confirmation booklet
QCDA (2010) December 2009 Standards confirmation presentation
QCDA (2010) December 2009 – Standards confirmation booklet
QCDA (2010) December 2009 – Standards confirmation minutes

Research reports commissioned by QCDA

Bramley, T. (2009) *Conceptualisation the reliability of single level tests*. A report to the NAA. Cambridge Assessment

Cito (2009) *Alternative methods of obtaining item statistics for single level tests: embedding items*.

Johnson, S. and Johnson, R. (2008) *Conceptualisation, establishing and monitoring of the reliability of single-level tests*. Assessment Europe Ltd

Pollitt, A., Derrieck, K. and Lynch, D. (2010) *Single level tests of KS2 writing: paired comparisons research report*. TAG developments

Pollitt, P. and Ahmed, A. (2009) *Conceptualisation of level 6 performance in the context of the key stage 3 programme of study*. Cambridge Exam Research.

Verstarlen, H. and Bechger, T. (2008) *Classification accuracy of educational tests*. Cito.

Warwick, I., Dickenson, M. and Worthen, L. (2009) *Report on conceptualisation of Level 6 performance in the context of the key stage 2 programme of study*. London Gifted and Talented.

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2011

This publication is also available on our website at www.ofqual.gov.uk

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346