National Foundation
for Educational Research

Partial Estimates of Reliability: Parallel Form Reliability in the Key Stage 2 Science Tests

Final Report

Sarah Maughan
Ben Styles
Yin Lin
Catherine Kirkup

September 2009

Ofqual This report has been commissioned by the Office of the Qualifications and Examinations Regulator.

Partial Estimates of Reliability: Parallel Form Reliability in the Key Stage 2 Science Tests

Partial Estimates of Reliability:

Parallel Form Reliability in the Key Stage 2 Science Tests

National Foundation for Educational Research

September 2009

OFQUAL/09/4553

Project Team

Sarah Maughan	Project Director
Catherine Kirkup	Project Leader
Elizabeth Maher	Project Administration Assistant
Ben Styles	Statistician
Yin Lin	Statistician

Contents

In May 2008, The Office of the Qualifications and Examinations Regulator (Ofqual) launched its Reliability Programme, which aims to stimulate debate about the reliability of assessments and to generate evidence about the extent of error in test and examination results in England. This report provides the results from a research project commissioned as part of Ofqual's Programme, which investigated the parallel form reliability of the key stage 2 science tests. These tests are taken each year by all pupils in year 6 (age 11) and each year a subset of pupils takes an equivalent test, which has been developed for use as the following year's live test. The levels that the pupils were awarded on each of these two versions of the tests were compared using a variety of statistical methods and the internal reliability of the tests was also calculated. Results from the analyses indicate that the tests have reasonably high internal consistency for tests of this nature and that the different forms of the test are reasonably parallel. Classification consistencies of 79 percent were found for the tests developed for each of the most recent three years, equivalent to a classification correctness of approximately 88 percent. These results are briefly compared to the results from similar analyses for the key stage 2 English tests.

1 Introduction

In May 2008, The Office of the Qualifications and Examinations Regulator (Ofqual) launched its Reliability Programme. This is a two year programme which aims to stimulate debate about reliability of assessments, and to generate research evidence that will provide a clearer picture of the magnitude of error in different tests and examinations. As part of this programme the National Foundation for Educational Research (NFER) has been commissioned to conduct a project to quantify the likelihood of pupils receiving a different national curriculum level if they sat a different version of the key stage 2 science test, that is, to conduct an investigation of the parallel form reliability of the tests.

Similar research has already been conducted for the key stage 2 English tests (NFER 2007, Newton 2009). The impetus for Newton's research came largely from an on-going debate, both in the media and within the educational community, concerning the extent of misclassification within national curriculum testing (Newton, 2009). In this context misclassification is used to mean cases where a pupil is awarded a level from their national curriculum test that is incorrect based on their true ability. The media debate was originally sparked by a claim by Professor Dylan Wiliam that at least 30 percent of pupils were likely to have been awarded an incorrect national curriculum level at key stage 2 (Wiliam, 2001). In order to add to the body of evidence on this issue, the same methodology was used for this study as for Newton's English studies, so that a direct comparison can be made between the results for the two sets of tests.

Section 2 of this report describes the methodology that has been used to analyse the data from the key stage 2 science tests, section 3 details the results from the analyses, section 4 discusses the results and compares these to the results for the key stage 2 English tests, and section 5 provides the concluding remarks.

2 Methodology

2.1 Data sets

The key stage 2 science test assesses levels 3 – 5 of the national curriculum and consists of two papers: paper A and paper B. Papers A and B have 40 marks each, giving a total of 80 marks. Pupils' marks from both papers are aggregated to calculate their overall science level. The test papers each have a time allowance of 45 minutes and are equivalent in terms of the curriculum areas and skills assessed.

The development cycle for a key stage 2 science test spans a three-year period and consists of a number of different stages, including a first pre-test in which the items are tested to ensure they function appropriately, and a second pre-test, which uses a final version of the papers to set the level thresholds for the new version of the test. The same test that is used in the second pre-test in year *x* will be used as the live test in year *x*+1[1].

The datasets used for this analysis are summarised in Table 1. For the purposes of this project, data that was collected as part of the second pre-test was compared with live test data for the same group of pupils[2,3]. The data also includes results from an anchor test that is administered alongside the pre-test papers. This anchor test is parallel in format to one of the final papers, and is used for statistically carrying forward the standards of levels from one year's set of test papers to another; in technical terms it is used for equating different versions of the test. The same anchor test is used over a number of years. The basic statistics for each of the tests used in the analyses are presented in Appendix 1.

The basic statistics provide an idea of the spread of pupil marks across the whole mark range. This is of interest in an analysis of this nature because features such as large numbers of pupils at particular marks, especially if these coincide with the cut scores could impact on classification consistency, for example if a level threshold changed by one mark there could be a large impact on classification error if a large percentage of pupils changed a level. Similarly, the spread of the cut scores, i.e. how many marks between them, could also have an impact, as it would be more likely for pupils to be misclassified by more than one level if the cut scores were only a few marks apart.

Each pre-test paper or anchor test was taken by a group of pupils in year 6, approximately four to six weeks prior to the live test window at the beginning of May[4]. This ensures pupils are as close as possible in age to pupils who will sit the

[1] In an ideal world no changes would be made to the tests after the second pre-test, but in reality some small changes are often made. Over recent years fewer changes have been made to the tests after the second pre-test.
[2] We used the live test data prior to any re-marking or appeals. This version of the data was selected as it was felt to be most similar to the pre-test data (for which there are no re-marks or appeals).
[3] Live test data is from the QCA dataset.
[4] The timing of the pre-test recently changed from March to April.

live test one year later. The sample of pupils used for the second pre-test was selected to be representative of the overall year 6 cohort. The pre-tests and anchor tests were administered in schools in as close to live test conditions as possible. The live test was then administered four to six weeks later to all pupils at key stage 2, including the pupils who had taken the pre-test of the following year's test earlier.

Analyses were conducted on the combined pre-test papers and the combined live test papers, as well as on the individual components of the pre-tests and the anchor test. For the purpose of quantifying the number of pupils who would have received a different level on a parallel version of the test, the first of these analyses would be sufficient. However, these tests have been administered in different conditions (see the discussion on the pre-test effects in section 2.2 below) so the analyses between the pre-test papers and with the anchor test have been included as a baseline against which the main results can be compared. No comparisons were made between papers A and B on the live tests as no item level data was available.

There was no anchor test administered in 2004. For the years 2005-06, 2006-07, 2007-08 and 2008-09, pairs of the pre-test A, pre-test B and anchor test were administered to different groups of pupils during pre-test 2. There were two rounds of pre-tests administered in 2006 in order to provide a bridge between pre-2006, when pre-tests were held in March and post-2006, when pre-tests were held in April. The timing of the pre-test was moved to reduce the magnitude of the pre-test effect. In this analysis the data for the first round of pre-tests was used, as this mapped to data from previous years for the 2007 equating.

Table 1: Summary of datasets.

Year of comparison	Combinations	Sample size
2004-2005	2005 pre-test (A+B) & 2004 live test (A+B)	900
	2005 pre-test A & 2005 pre-test B	901
2005-2006	2006 pre-test (A+B) & 2005 live test (A+B)	573
	2006 pre-test A & 2006 pre-test B	578
	2006 pre-test A & anchor test	430
	2006 pre-test B & anchor test	422
2006-2007	2007 pre-test (A+B) & 2006 live test (A+B)	645
	2007 pre-test A & 2007 pre-test B	655
	2007 pre-test A & anchor test	240
	2007 pre-test B & anchor test	234
2007-2008	2008 pre-test (A+B) & 2007 live test (A+B)	518
	2008 pre-test A & 2008 pre-test B	521

Year of comparison	Combinations	Sample size
	2008 pre-test A & anchor test	364
	2008 pre-test B & anchor test	364
2008-2009	2009 pre-test (A+B)& 2008 live test (A+B)	450
	2009 pre-test A & 2009 pre-test B	528
	2009 pre-test A & anchor test	360
	2009 pre-test B & anchor test	334

2.2 Cut scores

The key stage 2 tests are designed to assess pupils working at levels 3 to 5 of the national curriculum. Those pupils who do not achieve the level 3 threshold in the tests are said to be 'below level 3'. A combination of data from the second pre-test, judgemental evidence from standard setting exercises (eg script scrutiny), and data about the performance of the population as a whole, is used to set cut scores. Test equating is a statistical process by which scores on one test are compared to scores on another test to assess the relative difficulty of the two tests. This process is used to compare each national curriculum test with previous tests alongside the other evidence, thereby ensuring that the standards required to achieve each level are maintained consistently from year to year.

The final cut scores are agreed by a level setting panel appointed by the Qualifications and Curriculum Development Agency (QCDA)[5].

Cut scores for levels 3, 4 and 5 are available on all live tests. They are provided in Table 2 below.

Table 2: Cut scores on live tests.

Year	Level 3	Level 4	Level 5
2004	21	39	61
2005	23	42	63
2006	21	40	62
2007	23	41	62
2008	22	41	64

Cut scores are not generally available for pre-test papers, although in order to award levels for the pre-test 2 data for the purposes of these analyses, it was necessary to agree cut scores so that a comparison of the levels awarded could be conducted. As

[5] For full details of the level setting and test equating processes see the QCDA website http://testsandexams.qcda.gov.uk/18977.aspx

part of this project cut scores on all full pre-tests (i.e. on pre-test papers A and B as a whole test) were obtained from the 2007 and 2009 Draft Level Setting Reports produced by NFER for the QCDA.

As described above the pre-test 2 papers for year *x* are the same as the live papers for year *x*+1, so in theory it should be possible to use the cut scores set for the live tests to award levels to pupils who had sat the equivalent tests as part of pre-test 2. However, the two tests have been taken in different conditions in that the pre-tests were sat in a low stakes context: the results were not of importance to the pupils or the teachers, whereas the live tests were much higher stakes[6]. In addition the pre-tests were taken approximately four to six weeks prior to the live tests, during which period extra learning and revision are likely to have taken place. These factors have been termed the 'pre-test effect' and have been shown to have an impact on the pupils' results. Cut scores come at quite different points on the distribution of scores when live test to pre-tests are compared, whereas for a pre-test to pre-test comparison they tend to come at similar points. For fair comparisons to be made between the performance of pupils on the pre-test and on the live tests, different cut scores must be used (see Pyle et al, 2009 for a discussion of the pre-test effects in the key stage 2 science tests).

In order to remove the pre-test effect from any comparisons between tests sat under pre-test conditions and tests sat under live test conditions, cut scores for the pre-tests were obtained from the live test cut scores for any given year by means of equipercentile equating. As it is the same pupils taking both tests, the pre-test cut scores obtained in this way are equivalent to the live test cut scores, taking into account different testing conditions and any differences in the pre-test effect at different points on the ability range.

The cut scores for the pre-tests are provided in Table 3.

[6] The key stage 2 tests are used for accountability purposes. The percentage of pupils awarded level 4 or above is reported and is used as a measure of school effectiveness.

Table 3: Cut scores on pre-tests (pre-test A and B as a whole test) based on pre-test to live test equating.

Year	Level 3	Level 4	Level 5
2005*	19	35	55
2006	16	28	50
2007	17	29	54
2008	15	29	55
2009	15	29	55

* Note: Items were moved around for the 2005 live test after the 2005 pre-test (taken in 2004) and some items were altered, as the test had proved too easy during the second pre-test. New, more difficult items were swapped in from the reserve test. For the purpose of our comparison, the pre-test before item swapping (and the corresponding cut scores) is used, because it was taken by the same group of pupils who took the 2004 live test or the anchor test. The 2005 pre-test therefore represents a test that never went live in its entirety.

For the purpose of this project we also needed to calculate cut scores for the separate pre-test papers to allow comparisons with these to be made. In order to obtain cut scores on these individual test papers, item response theory (IRT) true score equating was carried out to equate them to the full pre-test. This process involved generating a measure of pupil ability through a two parameter IRT model. Using item parameters from the same model an expected score on each item of the pre-test papers/anchor test was generated for each pupil. These expected scores were summed to give the 'true score' on each paper or pair of papers. This allowed cut scores to be mapped from a pair of papers to the individual pre-test papers and the anchor test.

For a single year of comparison, the pre-tests and anchor tests were taken at about the same time under pre-test conditions and, given that they are equivalent in terms of the curriculum areas and skills assessed, it is reasonable to assume that the pre-test effects on them are similar. Therefore, the tests could be equated directly without concerns about pre-test effects, and the cut scores obtained are equivalent taking into account the testing conditions. This means that they can be used without adjustment when comparing parallel tests, as in this project.

The equating results are shown in Table 4.

Table 4: Cut scores for the separate pre-test papers based on IRT true score equating.

Year of comparison	Paper	Level 3	Level 4	Level 5
2004-2005	Pre-test A	10	17	28
	Pre-test B	9	18	27
2005-2006	Pre-test A	9	15	25
	Pre-test B	7	13	25
	Anchor	6	11	22
2006-2007	Pre-test A	8	14	27
	Pre-test B	9	15	27
	Anchor	6	11	24
2007-2008	Pre-test A	7	15	29
	Pre-test B	8	14	26
	Anchor	6	12	24
2008-2009	Pre-test A	9	15	28
	Pre-test B	6	14	27
	Anchor	6	12	25

* Note: The anchor test cut scores, as obtained by IRT true score equating described above, vary between years. Key stage 2 science level setting has always been susceptible to slight movement since the source cut scores are those from the previous year's live test rather than from the anchor test. This is because the anchor test has always been used for linking tests year by year and has not undergone a formal standard setting exercise itself. The variation between 06/07, 07/08 and 08/09 is one score point; a good indication that standards are being maintained. The slight variation could be explained by negligible differences in the equating results that are translated into a difference of one score point by rounding. Between 05/06 and 06/07 there is a change of two score points, which cannot be due solely to rounding.

2.3 Analysis

A variety of different techniques have been used to analyse the parallel form reliability of the different tests. A description of these and what they aim to show is given below.

1 **Classification consistency cross-tabulation**: in this analysis we produce a straightforward table of the level that the pupil was awarded at pre-test 2 against the level that the pupil was awarded in their live test. The table gives a percentage of pupils who were awarded the same level, those who were awarded one level different, and so on. This is perhaps the key measure for quantifying the number of pupils who would have received a different level, had they been given a different version of their key stage 2 science test.

2 **Kappa Statistic:** this analysis provides a statistical measure of the agreement of the levels awarded on the two forms of the test. Summing the percentages of agreement from the classification consistency tables gives a measure that includes the agreement between levels that would occur by chance alone. The Kappa statistic measures the extent of agreement excluding the possibility of agreeing by chance.

3 **Correlation coefficients:** Test reliability is defined as a test's ability to measure test takers' true ability accurately. One way to measure the reliability of a test is through the use of another test which has the same construct as the existing test, i.e. a *parallel* test. Truly parallel tests have the same internal consistency. This can be measured by their score correlation coefficient when the two tests are taken under the same conditions by the same group of individuals. The internal consistency measures for the separate tests should be the same as the correlation coefficient between the two tests (Thissen and Wainer, 2001, p30).

4 **Cronbach's alpha:** This is a commonly used statistic in test development processes. It provides a measure of the internal consistency of the test by comparing how each item performs individually with how all the items on the test perform together. The value of Cronbach's alpha increases when the correlations between the items increase. Cronbach's alpha is generally reported by QCDA for the key stage tests as an indication of how reliable they are.

5 **Rank order correlations:** an alternative perspective of whether the results on the two tests are comparable is to look at the rank order of the pupils on each of the tests. A correlation between the two rank orders provides a measure of how similar the two tests are at ranking pupils. Changes to the rank order may suggest that the different items in the two tests are having a different impact on different pupils. Question types included or topics covered are relatively easier or more difficult in one form than the other for certain pupils.

6 **Un-attenuated correlations:** the un-attenuated correlation is the correlation taking into account internal inconsistency; it tells us what the correlation would be if it were possible to measure the scores on the two tests with perfect reliability.

The different analyses each provide a different measure of the reliability of the key stage 2 science test and the results are provided in section 3 below and discussed in section 4.

2.4 Item classification

The nature of the items in the tests is likely to affect the reliability of the tests overall. Each year the item types included in the tests are categorised for review purposes. The classifications are given below for the 2004 live tests (Table 5) and the 2005 to 2008 live tests (Table 6). Unfortunately the categories used for the classification changed between 2004 and 2005 making a direct comparison between the 2004 and the other tests more difficult, but the later tests were all categorised in the same way.

Table 5: Breakdown of marks by item type in 2004 test.

Question Type	Multiple Choice	True/ False	Order/ Match	Table Key	Diagram	Graph	Open Response	
							Short	Long
Paper A	9	1	1	5	2	1	11	10
Paper B	9	2	1	6	2	0	6	14

The items were classified on the basis of how the pupils answer the questions. For example, where the pupils fill in a table or key, those items have been classified in that category, where the pupils interpret or use a table or key, those items would usually be classified as either open response or multiple choice items.

Short open response questions are those where only one or two words are required for an answer. Items that require the minimum of a short sentence are classified as long open response (e.g. Which factor would you change?).

Table 6: Breakdown of marks by item type in 2005, 2006, 2007 and 2008 tests.

	Closed response	Single word response	Open response	Total marks
2005				
A	14	14	12	40
B	13	10	17	40
	27	24	29	80
2006				
A	12	15	13	40
B	16	7	17	40
	28	22	30	80
2007				
A	14	11	15	40
B	11	13	16	40
	25	24	31	80
2008				
A	17	4	19	40
B	13	11	16	40
	30	15	35	80

For the purpose of this table, 'closed' response items include multiple choice, matching and true/false/can't tell items, and 'single word' responses include both one

word open response items and items where the answer can be found from a table/data provided. Open response items are those requiring pupils to write a phrase in order to demonstrate scientific understanding/knowledge, in an explanation for example.

2.5 Limitations of the Methodology

The analysis conducted here makes a number of assumptions about the data. Perhaps the most significant is the pre-test effect, and that the methods adopted have appropriately adjusted for any differences in performance based on the stakes and the timing of the administration of the tests. The method employed here for deriving cut scores for the pre-tests does take account of differences in the pre-test effect at different points on the ability range (see section 2.2), however, it does not take account of differences in item types. In a recent study it was found that short open response items exhibit a greater pre-test effect than closed items (see Pyle *et al*, 2009). However, the proportion of items of different types does not vary to any large extent from year to year. Also, if they did, the adjustment for the pre-test effect that is inherent in the equating method used would take this into account since it is based on how pupils perform in the test as a whole.

A second significant assumption is that any changes made after the second pre-test do not have an impact on the comparisons made between the sets of data. This assumption is possible because such changes, if any, are usually kept to a minimum, and where such changes affect marking/scoring, these are taken into account for equating.

Clearly the cut scores themselves are crucial to any of the measures in this report that concern misclassification. Since they were established using equating procedures, they are optimal cut scores for the maintenance of standards between tests and papers. Any departure from the cut scores used here is likely to lead to an increase in misclassification. The implications of such changes would be an interesting area for further reliability research.

The basic statistics provided in Appendix 1 for each of the tests used in these analyses suggest that, in general, the majority of the mark range is used, with few pupils getting less than 10 marks on the tests, but pupils achieving 78, 79 or 80 out of the 80 mark test in most tests. The live tests generally have a mean mark of between 55 and 60 marks and a standard deviation of about 13 marks. This would suggest that there is some skew in the results towards the higher marks, but the pupils are spread out over most of the range. These statistics showing that most of the range is used may suggest that bunching of pupils around particular marks has been kept to a minimum, thereby minimising any effect from the cut scores. It may be, however, that by making the test a little harder, so the bottom 10 marks are also used effectively, there would be a small improvement in reliability related to cut score effects.

There are a number of other limitations to the analysis that is being conducted here. Newton (2009) explains that there are two broad categories of causes by which pupils could be assigned an inaccurate classification: random or systematic. Systematic

errors are those that are inherent within the process or system and would therefore be replicated for a pupil if the assessment were repeated. These would not be picked up by these analyses. Random errors are '*unsystematic causes of error in assessment results*'. '*It concerns the likelihood that students would have received different results if they happened to have been allocated a different test date, a different version of the test, a different marker and so on'*. This analysis focuses on the extent of error associated with awarding national curriculum levels using two different versions of the same test, in other words Newton's '*different version of the test'*. However, as the test is not sat at an identical time, it is not possible to disentangle other possible causes of error, such as Newton's '*different test date'*.

The analysis is examining the reliability of supposedly parallel tests by means of classification consistency. Differences in classification found in our analyses could arise for a number of reasons:

- the tests, although very similar, might not be completely equivalent in construct and could be measuring slightly different content or skills, leading to slightly different ability sets being tested;

- variation in pupil performance due to factors unrelated to the tests themselves, for example, if a boundary level 4 pupil took two truly parallel tests, it is still possible that the two tests would produce different level classifications, as a result of variation in individual performance (for example the pupil may be suffering from hay fever on one test session and not on the other);

- changes in the way scripts were marked between the two papers, as part of the pre-test or as part of the live test marking may have had an impact on the marks awarded;

- internal inconsistency of the tests.

The study did not aim to differentiate the relative contributions to classification consistency of the different sources.

3 Results

3.1 Classification consistency crosstabs

Cut scores as identified in the previous section were used to award levels to pupils on each test: the pre-test papers, the anchor test and the live test as relevant. The levels from the different versions of the test have been used to produce a cross-tabulation of results for different pupils. In each comparison, the live test level is the 'real' level as awarded and reported, the pre-test or anchor test level is an approximate level using cut scores set for the purposes of this project and tests sat in pre-test rather than live test conditions.

For each pair of tests, a cross-tabulation gives direct presentation of any differences in classification between the two tests (see Appendix 2 for all the tables).

The results of most interest here are those which compare levels attained on the pre-test (A + B) when compared to the live test (A + B) since these refer to the complete test. This analysis compares the levels awarded on the live test with the parallel version of the test as used in the pre-test. The percentages of pupils who were awarded the same level on each version of the test were added up (e.g. for the 2005 pre-test comparison with the 2004 live test:

0.4% (<L3) + 7.6% (L3) + 29% (L4) + 34.8% (L5) = 71.8%).

In other words, we summed the percentages of pupils who fell on the diagonals of each table.

The percentage of agreement in each of the years is given below:

2005 pre-test v 2004 live test	72%
2006 pre-test v 2005 live test	74%
2007 pre-test v 2006 live test	79%
2008 pre-test v 2007 live test	79%
2009 pre-test v 2008 live test	79%.

There would appear to have been an improvement in the classification consistency of the tests over the five years, with the last three years being better than the first two.

Almost all of the remainder of the pupils were classified into the adjacent level, with less than 1 percent of pupils being awarded more than one level different in four of the five years (in the other year no pupils were awarded more than one level different).

3.2 Kappa statistics

Summing the percentages of agreement from the classification consistency tables gives a measure that includes the agreement between levels that would occur by chance alone. The Kappa statistic measures the extent of agreement excluding the possibility of agreeing by chance. The extent of agreement by chance will vary according to how many levels/grades are assigned. When comparing the classification consistency of assessments that have different numbers of levels/grades, Kappa would be particularly useful, although this does not apply here. For each comparison, the consistency from the classification crosstabs is presented alongside Kappa in Table 7.

Table 7: Kappa statistic and level of consistency.

Year of comparison	Papers	Kappa	Consistency (%)
2004/2005	Pre-test A+B vs live test (A+B)	0.54	72
	Pre-test A vs Pre-test B	0.45	66
2005/2006	Pre-test A+B vs live test (A+B)	0.55	74
	Pre-test A vs Pre-test B	0.55	74
	Pre-test A vs anchor test	0.58	74
	Pre-test B vs anchor test	0.58	76
2006/2007	Pre-test A+B vs live test (A+B)	0.63	79
	Pre-test A vs Pre-test B	0.53	73
	Pre-test A vs anchor test	0.68	82
	Pre-test B vs anchor test	0.49	71
2007/2008	Pre-test A+B vs live test (A+B)	0.61	79
	Pre-test A vs Pre-test B	0.60	78
	Pre-test A vs anchor test	0.58	76
	Pre-test B vs anchor test	0.62	77
2008/2009	Pre-test A+B vs live test (A+B)	0.64	79
	Pre-test A vs Pre-test B	0.61	77
	Pre-test A vs anchor test	0.58	75
	Pre-test B vs anchor test	0.58	75

It can be seen from these results that all the tests account for a large degree of similarity in results between pupils after chance is taken into account. This is true for comparisons between the live test and the parallel version of the test used in the pre-test, and between the pre-test papers themselves and the anchor tests. The Kappa statistics improve over time, with over 0.6 agreement in the latter three years, coinciding with the improved classification consistency that is seen in those years.

3.3 Correlation coefficients and Cronbach's alpha

For this study, the correlation coefficient was computed for pairs of assumed parallel tests. For each year of comparison, pre-test A, pre-test B and the anchor test have very similar constructs and, for the purpose of this work, were assumed to be parallel. Similarly, the live test and the pre-test as whole tests (each comprising papers A and B) were also assumed to be parallel tests. The following table (Table 8) documents the raw score correlation coefficients, Cronbach's alpha for each of the two tests in question, the un-attenuated correlation and the rank order correlation. Cronbach's alpha has not been computed for live test papers since item level data was not available, although where given this has been taken from the test statistics published on the QCDA website.

Table 8: Raw score correlation coefficients, Cronbach's alpha, un-attenuated correlations and rank order correlations.

Year of comparison	Papers	Correlation	Cronbach's alpha	Un-attenuated correlation	Rank order correlation
04/05	Pre-test A+B vs live test (A+B)	0.85	0.92/*	(requires alpha on both tests)	0.85
	Pre-test A vs Pre-test B	0.82	0.88/0.85	0.95	0.81
05/06	Pre-test A+B vs live test (A+B)	0.81	0.93/0.92*	0.88	0.79
	Pre-test A vs Pre-test B	0.84	0.85/0.88	0.97	0.83
	Pre-test A vs anchor test	0.86	0.87/0.88	0.98	0.85
	Pre-test B vs anchor test	0.86	0.86/0.87	0.99	0.86
06/07	Pre-test A+B vs live test (A+B)	0.85	0.92/0.93*	0.92	0.84
	Pre-test A vs Pre-test B	0.83	0.85/0.84	0.97	0.81
	Pre-test A vs anchor test	0.87	0.86/0.88	1.00	0.87
	Pre-test B vs anchor test	0.78	0.86/0.88	0.90	0.81
07/08	Pre-test A+B vs live test (A+B)	0.86	0.94/*	(requires alpha on both tests)	0.85
	Pre-test A vs Pre-test B	0.86	0.89/0.89	0.97	0.84
	Pre-test A vs anchor test	0.85	0.87/0.87	0.97	0.84
	Pre-test B vs anchor test	0.88	0.89/0.88	1.00	0.89
08/09	Pre-test A+B vs live test (A+B)	0.88	0.94/*	(requires alpha on both tests)	0.88
	Pre-test A vs Pre-test B	0.85	0.86/0.90	0.97	0.84
	Pre-test A vs anchor test	0.85	0.86/0.88	0.98	0.84
	Pre-test B vs anchor test	0.88	0.90/0.90	0.98	0.89

***** Cronbach's alpha has not been computed for live test papers since item level data was not available. Where given for the 2005 and 2006 live tests this has been taken from the test statistics published on the QDCA website.

It can be seen that the results for the pre-test A+B and live test A+B analyses are largely similar to the pre-test A and pre-test B comparisons. This would suggest that the papers used in consecutive years are as similar as the two papers used within one year. It can also be seen that the anchor paper tends to have similar correlations with each of the pre-test A and the pre-test B papers suggesting that the A and B versions are similar, or are both as similar to the anchor test as to the other.

Over time the correlations between the anchor test and the pre-test papers remain fairly similar (as mentioned above the same anchor test is used on each occasion) suggesting that the tests remain fairly similar over time.

In terms of whether the tests are reliable enough, it is stated on the QDCA website that '*Cronbach's alpha measures a test's reliability. Figures of less than 0.5 indicate a test which is likely to be unreliable; figures of 0.8 or higher are an indication of a reliable test'.* It can be seen from the table that the Cronbach's alpha for the combined pre-tests and for the live tests where these are available all exceed 0.9 making these very reliable tests against that criteria. As would be expected, the Cronbach's alpha for the individual pre-test papers and for the anchor test are lower (Cronbach's alpha is affected by the overall length of the test and the individual test papers are shorter than the combined tests), but these are still consistently above 0.8 and in most cases are above 0.85.

As described in the methodology section above, the Cronbach's alpha is a measure of the internal consistency of the tests. The correlation coefficients and the rank order correlations compare the pupil results across different tests. As might be expected these results are lower than the Cronbach's alpha. This difference across the different measures highlights the fact that the different forms of the test are not entirely parallel, and that there is greater similarity within a single year's tests, than there is between tests from different years. However, the un-attenuated correlations give the correlation taking into account internal consistency. The high values for the un-attenuated correlation coefficients indicate that the tests are largely parallel.

The results from the correlations seem fairly consistent over the years, and are fairly consistent irrespective of the comparison that is being made, whether it is live test to pre-test or between pre-tests.

4 Discussion

The results from a number of different analyses on the data from the key stage 2 science tests over the last five years are given above. The different analyses all provide a different measure of the reliability of the tests; assuming the different forms of the tests are parallel. The purpose of this is to quantify the degree of unreliability in the levels awarded from the tests based on different versions, or in other words, how many pupils would be awarded a different level if they had sat a different version of the test.

A key analysis in answering this question is the production of cross-tabulations of levels awarded on each of two versions of the test. These cross-tabulations show relatively high levels of pupils achieving the same level on the two tests, with 79 percent of pupils receiving the same level on the pre-test A + B papers and on the live tests in the last three years. However, these percentage differences are based on results in two tests. As Newton points out, these percentages are '*estimates of classification consistency, the extent to which results from the two tests agree. They are not estimates of classification 'correctness''* (Newton, 2009, p.201). It is important to note that consistency does not mean that the level awarded is necessarily correct - a pupil may have been awarded the 'incorrect' level twice. Equally, inconsistency does not mean that neither level is correct.

Classification 'correctness' is the probability that a pupil is awarded the 'correct' level on the basis of one of the two test administrations and where 'correctness' is taken to be correspondence with a pupil's 'true' level. Newton goes on to say that classification correctness is therefore likely to be higher than classification consistency. Using his equation for the probability of being classified inconsistently: $2p(1-p)$, where p is the probability of being classified correctly, a further analysis was conducted to quantify the differences that would be seen. The results from this analysis are shown below.

2005 pre-test v 2004 live test	$2p\ (1-p) = 0.28,\ p = 0.83$
2006 pre-test v 2005 live test	$2p\ (1-p) = 0.26,\ p = 0.85$
2007 pre-test v 2006 live test	$2p\ (1-p) = 0.21,\ p = 0.88$
2008 pre-test v 2007 live test	$2p(\ 1-p) = 0.21,\ p = 0.88$
2009 pre-test v 2008 live test	$2p\ (1-p) = 0.21,\ p = 0.88.$

These results show that between 83 percent and 88 percent of pupils would be likely to be given the correct level when sitting the tests described in this study.

The tests over the last three years appear to have become more reliable which may be expected as a number of changes have been introduced over recent years to improve the processes: there are more controls against making changes to the test after pre-test 2, and there is now a tendency to use the reserve papers from the year before as a starting point for the new set of papers, rather than starting from scratch. This means

that the questions included will frequently have had one additional pre-test during which they could be polished.

In section 2, a breakdown of the number of items of different types included in the tests is given. This classified items as closed response, single word response and open response in most cases. The tables showed some variation in the item types included, however, the classification consistencies achieved with the different tests are remarkably stable, suggesting that the inclusion of items of different types is not having a huge impact on the variability of the levels awarded.

The variation in item types could be expected to have an impact on marking reliability, which may in turn increase the level of classification inconsistency. The findings here would suggest that marking reliability is not having a huge impact, although a further study which involved specific research is needed in order to really quantify the effects.

The classification consistency figures for key stage 2 science (ranging from 72 percent to 79 percent for the pre-test A + B to live test comparison) compare well with the figures for key stage 2 English which Newton (2009) reports. Newton's data covers one year only, the 2006 pre-test and the 2005 live test, and is given for reading and writing separately and for English as a whole. For these tests the classification consistencies are: reading, 73 percent, writing 67 percent and English as a whole, 73 percent. The higher degree of classification consistency in the science tests is likely to be due to the more objective nature of the items.

The classification consistency of the key stage 2 English tests can be converted in the same way as described for the science tests above, to a classification 'correctness' of 84 percent. Again, the figures for key stage 2 science, with a classification correctness of 88 percent in the last three years, compare favourably with this figure. These figures for the science and the English results can be transformed into a measure of classification error by subtracting the classification correctness figures from 100, so for key stage 2 science we get 100 – 88 = 12. These figures can then be compared with the 30 percent of pupils receiving an incorrect level suggested by Wiliam in 2001. It can be seen that the results from the analyses reported here are much better than the results suggested by Wiliam. It is not clear why this degree of difference arises, although his results were produced using a different methodology and mathematical simulations. An Ofqual seminar planned for later in 2009 will aim to investigate these differences more fully.

A number of the pupils would, clearly, be awarded the same level on a test purely by chance, and the Kappa statistic was used to quantify the proportion of pupils who would receive the same level once the effects of chance are removed. The figures from this analysis were lower than from the classification consistency measures described above, as would be expected. However, these analyses also suggest that the key stage 2 science tests are reasonably accurate at assigning the correct level to pupils. For the Cronbach's alpha results QCDA have published guidelines about what

is an acceptable result for tests of this type. Unfortunately no such guidelines are available for the Kappa statistic.

Based on simple cross-tabulations and Kappa statistics, it is hard to address how much of the classification difference was accounted for by each of the reasons stated in section 2: variability in the tests themselves, variability in the performance of the pupils, or unreliability in individual tests. A number of further analyses were conducted in order to investigate some of these factors in more detail.

Cronbach's alpha statistics were produced for the pre-test versions and the anchor tests, and were taken from the QCDA website for the live tests (where these are published). All of the results for the pre-test versions and for the live tests were in excess of 0.9. The acceptable standard for tests such as these is 0.8, so it is clear that the key stage 2 science tests are very reliable when compared to this measure. This suggests that only a limited amount of the variability in the levels awarded could be accounted for by unreliability in individual tests.

The results of the correlations versus Cronbach's alpha comparisons call in to question the assumption that the test papers used for the analysis are truly parallel. Cronbach's alpha is a lower bound for a test's reliability (Thissen and Wainer 2001, p33). Since it is often higher than the correlation between our supposedly parallel test forms, this suggests that the two test forms are not, in fact, entirely parallel. Two truly parallel tests have the same expected score for each individual (equal to the true score). They also have the same observed score variance across individuals. If the questions in one test are not measuring exactly the same underlying construct as those in the other, they will not be parallel. This is very likely to be the case for two key stage 2 science papers/tests so these results are not particularly surprising.

One way of assessing the degree to which the two papers on any given test are parallel is to apply the Spearman-Brown adjustment for length to Cronbach's alpha on each paper. This allows alpha to be projected for a situation where a test is doubled in length in an exactly parallel way. By comparing projected alphas with those calculated for two pre-test papers together, we get an indication of how appropriate our assumption of the papers being parallel is. These figures are shown in Table 9.

Table 9: Actual and projected values for Cronbach's alpha.

Year of comparison	Paper	Cronbach's alpha	Cronbach's alpha projected for an 80 mark test	Actual value of alpha for papers A+B
04/05	A	0.88	0.93	0.92
	B	0.85	0.92	
05/06	A	0.85	0.92	0.93
	B	0.88	0.94	
06/07	A	0.85	0.92	0.92
	B	0.84	0.92	
07/08	A	0.89	0.94	0.94
	B	0.89	0.94	
08/09	A	0.86	0.93	0.94
	B	0.90	0.95	

Since the projected alphas correspond well with the actual values, we can use this as evidence that the two papers are consistent in reliability, that they are psychometrically close to being parallel and that the various samples of pupils used were very similar. Further evidence of the parallel nature of tests whose correlations were calculated is present in Table 8; the high values for the un-attenuated correlation coefficients indicate that the tests are highly parallel.

The correlations between pre-tests and live tests, and between the individual pre-test papers and the anchor tests are stable over time suggesting that the tests are maintaining a well-established pattern. For tests such as the key stage 2 science tests, which aim to sample content and skills from across the whole of the key stage 2 science curriculum, with different areas being covered in different years, it would be expected that differences would be seen between the tests, but these differences appear to be consistent over time. Two further measures of the reliability of the tests in different forms are the correlation coefficients and the rank order correlations. These two analyses provide very similar results irrespective of the test being compared, and are fairly consistent over time.

It is useful to be able to generate measures of misclassification using data collected from individual tests; this has obvious applications in the development of new tests. The conversion of Cronbach's alpha into a measure of misclassification is possible, although it has two problems associated with it: it does not easily take into account regression to the mean and it assumes the same standard error of measurement across the score range. Regression to the mean implies that the true score distribution around any given score point away from the mean score is skewed, and not centred around the observed score. Furthermore, we cannot assume a normal distribution for true scores given observed scores. Finally, test items being designed predominantly to discriminate between individuals towards the middle of the ability distribution implies

that the standard error of measurement increases as we move towards the ends of the score range. Within Classical Test Theory, methods have been devised to address these issues but an alternative way of approaching the problem is to use IRT.

IRT decision accuracy can be used to predict levels of misclassification during test development. Using this approach, regression to the mean is inherent to the calculation of true score distributions around each observed score and variable standard error of measurement is observed across the score range. It relies on the various assumptions of IRT not being violated, however, IRT assumptions have been tested for key stage 2 science and have been shown to be adequately satisfied for the use of the two-parameter model. The results of these tests are reported in recent key stage 2 science Draft Level Setting reports.

IRT decision accuracy and consistency calculations have been conducted for the 2009 pre-test as an illustration and the results are presented in the box below. We propose that IRT decision accuracy is presented alongside all equating information in future national test development work. This would provide easily understandable information on expected misclassification at the time of developing the test and would feed back into any decisions about how to structure the test to ensure high levels of correct classification. However, it is acknowledged that this approach would translate internal test inconsistency into decision accuracy. It would not incorporate other aspects of parallel-form reliability; Generalisability Theory may be worth exploring for this.

Calculating Decision Accuracy and Decision Consistency Using IRT

The IRT definition of true score is the expected value of a pupil's score given their ability. Using this definition, decision accuracy can be calculated from item parameters of an IRT model:

Decision accuracy at each score point is defined as the probability that a pupil's true score would be awarded the same level or grade as their observed score. Another way of saying this is that it is the probability that the grade or level awarded to a pupil is "correct". For an individual cut score this would be the probability that a pupil with a given observed score has a true score the same side of a specified cut score.

Similarly, and in direct comparison to the results reported here, IRT item parameters can be used to calculate decision consistency:

Decision consistency is defined as the probability that a pupil would be awarded the same level or grade on another parallel test as they were awarded on this test. (A parallel test is defined as one consisting of items with the same parameters as the test in question.) For an individual cut score this would be the probability that a pupil would get a result the same side of a specified cut score in another test.

The following results show the decision accuracy and decision consistency for each cut score on pre-test A + B 2009 using IRT model item parameters:

Decision accuracy: L3 100%; L4 97%; L5 92%; Overall 89%

Decision consistency: L3 99%; L4 96%; L5 89%; Overall 84%

The decision accuracy for the L4 cut score means 97 percent of pupils have a true score that is the same side of this threshold as their observed score. These individual percentages are hence determined both by the reliability of the test at each threshold and the proportion of individuals at each threshold. Since so few pupils score around L3, decision accuracy is very high at this point (99.53%). The overall accuracy provides an estimate that 89 percent of pupils are awarded the correct level and reflects the summation of decision inaccuracy at each level. This is comparable to the correctness measure (Newton, 2009, p.201) calculated as 88 percent using parallel test forms for this study.

The decision consistency for the L4 cut score suggests that 96 percent of pupils who have a score above or below this cut score would get a score that is the same side of this threshold in another parallel test. The overall consistency suggests that 84 percent of pupils would be awarded the same level if they took another parallel test. This is comparable to the figure of 79 percent classification consistency obtained using actual parallel test forms for this study. These figures are different since both approaches rely on various assumptions which are never completely satisfied. Further work would be needed to attempt to explain this difference; some of it may be because the parallel form work captures different aspects of error variance.

The various analyses that have been conducted have quantified the classification consistency and the classification correctness of the key stage 2 science tests. They have also investigated the impact of differences in reliability of the individual tests included in the comparisons and gone some way towards investigating the impact of different item types on the classification consistency, which could be considered as a loose proxy for the effects of marking reliability. However, no investigations have provided evidence on the impact of pupil differences on the classification consistency of the tests, and this is an area that could warrant further research. This could involve reviewing the performance of pupils at the borderlines and investigating the levels awarded and strengths and weaknesses in different areas of the curriculum. Further work would also be useful in the area of marking reliability and in the use of IRT for producing a measure of decision accuracy and decision consistency.

5 Concluding Remarks

The analyses conducted as part of this research have demonstrated that there is some limited variation between the different versions of the key stage 2 science test over the five years considered. Differences have been highlighted in the item types included and in the higher Cronbach's alpha results (measuring internal consistency of the tests) as compared to the correlation coefficients (measuring the similarity between different versions of the tests). This is what would be expected of curriculum-based tests like the key stage 2 science tests.

However, the tests themselves have very high levels of internal reliability and the different versions of the test perform in very similar ways, suggesting that the model for the tests is being maintained consistently over time.

In terms of a quantification of the parallel forms reliability of the tests, or the answer to the question, how many pupils would receive a different level if they were given a different version of the test, reasonably high levels of classification consistency are found, especially in the last three years, where a classification consistency of 79 per cent is seen.

It is difficult to draw conclusions about how this compares with other tests, such as whether tests with different item types, tests of different lengths or tests for different subject areas, would have vastly different results, as there is limited quantified data available. However, Newton (2009) did publish similar results for key stage 2 English tests. The results for English showed a classification consistency of 73 percent in the one year studied. The results for the key stage 2 science tests compare favourably with these results and therefore one could conclude that these tests achieve as high a value of consistency as might be expected from tests of this type. Whether this level of consistency is satisfactory is a matter of policy.

However, it should be noted that although this study produces similar findings to those produced by Newton, the context for both pieces of work was national curriculum testing for pupils at the end of primary education. It is not possible to generalise these results to other forms of assessment where the item types used, the number of grade boundaries, or the score distributions might be significantly different. Further investigation using data from different types of tests and examinations, whilst still using the methods of measuring the uncertainty of test results used in this report, would be valuable. This would help to build up a picture of the range of results found that could then be translated into policy decisions about what level of classification consistency or classification correctness is acceptable. Similarly, further investigation into different methods for measuring the uncertainty of test results, for example using decision accuracy and decision consistency, could build upon the work in this area.

A number of different methods have been used in this project in an attempt to quantify the parallel form reliability of the tests. When answering the question, would the pupils receive the same level if they had sat a different version of the test?, the

classification consistency crosstabs provide the most useful results. Although a number of assumptions need to be made about factors such as the pre-test effect, this analysis comes closest to providing evidence of levels awarded to pupils sitting different versions. However, in reality, it is not always possible to have different versions of the test taken by different pupils, so the measures of decision accuracy and decision consistency from the IRT analyses, which can be calculated from a single test, can provide a useful substitute that can be used as part of a test development process.

6 References

National Foundation for Educational Research (2007). *Submission to the Education and Skills Committee inquiry into testing and assessment.* Submission to the Education and Skills Committee Inquiry into Testing and Assessment. London: Qualifications and Curriculum Authority.

Newton, P. E. (2009). The reliability of results from national curriculum testing in England, *Educational Research*, **51**, 2, 181-212.

Pyle, K., Jones, E., Williams, C. and Morrison, J. (2009).Investigation of the factors affecting the pre-test effect in national curriculum science assessment development in England, *Educational Research*, **51**, 2, 269-282.

Thissen, D. and Wainer, H. (2001) *Test Scoring* Mahwah, New Jersey: IEA

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

Appendix 1

Year of comparison	Papers	Mean	Standard deviation	Minimum	Maximum	N
04/05	Pre-test A+B	50.9	13.7	9	78	900
	Live test (A+B)	55.0	13.3	13	78	900
	Pre-test A	25.7	7.4	5	40	901
	Pre-test B	25.2	7.0	4	40	901
05/06	Pre-test A+B	49.1	14.2	7	78	573
	Live test (A+B)	60.0	12.3	14	79	573
	Pre-test A	24.5	7.1	5	40	578
	Pre-test B	24.4	7.8	2	40	578
	Anchor	20.1	8.0	1	37	430
06/07	Pre-test A+B	48.9	13.1	10	77	645
	Live test (A+B)	57.4	12.4	13	79	645
	Pre-test A	23.9	7.1	3	39	655
	Pre-test B	24.8	6.8	5	38	655
	Anchor	22.6	7.7	4	39	240
07/08	Pre-test A+B	54.0	14.5	9	80	518
	Live test (A+B)	61.0	11.7	18	80	518
	Pre-test A	27.8	7.7	2	40	521
	Pre-test B	26.0	7.6	2	40	521
	Anchor	23.4	7.7	3	38	364
08/09	Pre-test A+B	50.2	15.4	10	77	450
	Live test (A+B)	58.9	13.6	12	80	450
	Pre-test A	25.9	7.4	6	39	528
	Pre-test B	24.7	8.3	2	40	528
	Anchor	22.4	7.9	4	38	360

For pre-test A+B and live test (A+B), score statistics were computed for pupils with valid scores on both tests. For individual papers A and B, score statistics were computed for pupils with valid scores on both papers. For the anchor test, score statistics were computed for pupils with valid scores on paper A and the anchor. See section 2.1 for an explanation of the datasets.

Appendix 2

Table A2.1: Differences in the classification of pupils between 2005 Pre-test A+B v 2004 Live Test A+B

		2004 Live test			
		Below L3	L3	L4	L5
2005 Pre-test AB	Below L3	<1	1	0	0
	L3	<1	8	4	<1
	L4	<1	4	29	9
	L5	0	0	9	35

Due to rounding, percentages may not sum to 100

Table A2.2: Differences in the classification of pupils between2006 Pre-test A+B 2005 Live Test A+B

		2005 Live test			
		Below L3	L3	L4	L5
2006 Pre-test AB	Below L3	<1	<1	0	0
	L3	<1	4	2	<1
	L4	0	3	28	9
	L5	0	0	10	42

Due to rounding, percentages may not sum to 100

Table A2.3: Differences in the classification of pupils between2007 Pre-test A+B v 2006 Live Test A+B

		2006 Live test			
		Below L3	L3	L4	L5
2007 Pre-test AB	Below L3	<1	<1	0	0
	L3	0	6	2	0
	L4	0	3	38	8
	L5	0	0	8	34

Due to rounding, percentages may not sum to 100

Table A2.4: Differences in the classification of pupils between2008 Pre-test A+B v 2007 Live Test A+B

		2007 Live test			
		Below L3	L3	L4	L5
2008 Pre-test AB	Below L3	0	<1	<1	0
	L3	<1	3	3	0
	L4	0	2	26	9
	L5	0	0	6	49

Due to rounding, percentages may not sum to 100

Table A2.5: Differences in the classification of pupils between 2009 Pre-test A+B v 2008 Live Test A+B

		2008 Live test			
		Below L3	L3	L4	L5
2009 Pre-test AB	Below L3	<1	<1	0	0
	L3	<1	6	2	0
	L4	0	4	38	7
	L5	0	0	7	35

Due to rounding, percentages may not sum to 100

Table A2.6: Differences in the classification of pupils between 2005 Pre-test A v 2005 Pre-test B

		2005 Pre-test B			
		Below L3	L3	L4	L5
2005 Pre-test A	Below L3	<1	1	<1	0
	L3	<1	7	4	0
	L4	0	5	23	12
	L5	0	0	10	36

Due to rounding, percentages may not sum to 100

Table A2.7: Differences in the classification of pupils between 2006 Pre-test A v 2006 Pre-test B

		2006 Pre-test B			
		Below L3	L3	L4	L5
2006 Pre-test A	Below L3	<1	<1	0	0
	L3	<1	4	3	0
	L4	<1	3	26	12
	L5	0	0	7	43

Due to rounding, percentages may not sum to 100

Table A2.8: Differences in the classification of pupils between 2007 Pre-test A v 2007 Pre-test B

		2007 Pre-test B			
		Below L3	L3	L4	L5
2007 Pre-test A	Below L3	<1	0	0	0
	L3	<1	4	4	0
	L4	<1	2	37	12
	L5	0	0	8	32

Due to rounding, percentages may not sum to 100

Table A2.9: Differences in the classification of pupils between 2008 Pre-test A v 2008 Pre-test B

		2008 Pre-test B			
		Below L3	L3	L4	L5
2008 Pre-test A	Below L3	<1	<1	0	0
	L3	1	4	2	0
	L4	<1	2	26	9
	L5	0	0	7	49

Due to rounding, percentages may not sum to 100

Table A2.10: Differences in the classification of pupils between 2009 Pre-test A v 2009 Pre-test B

		2009 Pre-test B			
		Below L3	L3	L4	L5
2009 Pre-test A	Below L3	<1	1	0	0
	L3	<1	5	3	0
	L4	<1	3	34	8
	L5	0	0	7	37

Due to rounding, percentages may not sum to 100

Table A2.11: Differences in the classification of pupils between 2006 Pre-test A v 2006 Anchor

		2006 Anchor			
		Below L3	L3	L4	L5
2006 Pre-test A	Below L3	2	1	0	0
	L3	2	4	4	0
	L4	0	4	31	8
	L5	0	0	7	38

Due to rounding, percentages may not sum to 100

Table A2.12: Differences in the classification of pupils between 2007 Pre-test A v 2007 Anchor

		Anchor 2007			
		Below L3	L3	L4	L5
2007 Pre-test A	Below L3	0	0	0	0
	L3	<1	4	0	0
	L4	<1	3	36	8
	L5	0	0	4	43

Due to rounding, percentages may not sum to 100

Table A2.13: Differences in the classification of pupils between 2008 Pre-test A v 2008 Anchor

		2008 Anchor			
		Below L3	L3	L4	L5
2008 Pre-test A	Below L3	<1	0	0	0
	L3	1	3	2	0
	L4	<1	4	30	11
	L5	0	0	6	43

Due to rounding, percentages may not sum to 100

Table A2.14: Differences in the classification of pupils between 2009 Pre-test A v 2009 Anchor

		2009 Anchor			
		Below L3	L3	L4	L5
2009 Pre-test A	Below L3	<1	<1	0	<1
	L3	<1	6	2	0
	L4	<1	5	31	7
	L5	0	0	9	38

Due to rounding, percentages may not sum to 100

Table A2.15: Differences in the classification of pupils between2006 Pre-test B v 2006 Anchor

		2006 Anchor			
		Below L3	L3	L4	L5
2006 Pre-test B	Below L3	<1	<1	0	0
	L3	1	3	3	0
	L4	0	5	36	8
	L5	0	<1	7	36

Due to rounding, percentages may not sum to 100

Table A2.16: Differences in the classification of pupils between Pre-test B 2007 v 2007 Anchor

		2007 Anchor			
		Below L3	L3	L4	L5
2007 Pre-test B	Below L3	2	<1	<1	0
	L3	<1	3	3	0
	L4	<1	<1	30	14
	L5	0	0	9	37

Due to rounding, percentages may not sum to 100

Table A2.17: Differences in the classification of pupils between Pre-test B 2008 v Anchor 2008

		2008 Anchor			
		Below L3	L3	L4	L5
2008 Pre-test B	Below L3	1	1	0	0
	L3	1	4	3	0
	L4	<1	4	33	4
	L5	0	0	9	39

Due to rounding, percentages may not sum to 100

Table A2.18: Differences in the classification of pupils between Pre-test B 2009 v Anchor 2009

		Anchor 2009			
		Below L3	L3	L4	L5
2009 Pre-test B	Below L-+3	<1	<1	0	0
	L3	<1	6	5	0
	L4	<1	6	27	7
	L5	0	0	8	41

Due to rounding, percentages may not sum to 100

First published by The Office of the Qualifications and Examinations Regulator in 2009.

Ofqual is part of the Qualifications and Curriculum Authority (QCA). QCA is an exempt charity under Schedule 2 of the Charities Act 1993.