London Review of Education
Vol. 3, No. 1, March 2005, pp. 29–45

Investigating the reliability of the Key Stage 2 test results for assessing individual pupil achievement and progress in England

Lesley Doyle[a]* and Ray Godfrey[b]
[a]*Stirling University, UK;* [b]*Canterbury Christ Church University College, UK*

'Personalised learning' and the value of national assessment data in achieving it have been identified by the UK Secretary of State for Education and Skills as essential for raising educational standards. Employing multilevel analysis, this paper compares children's end of primary school (Key Stage 2) test scores with those they achieved in comparable test papers taken in each term of their first year of secondary school. The paper questions the reliability of national assessment data in respect of the performance of individual children, their predictive validity and thus their value in contributing to the provision of 'personalised learning'.

Introduction

Speaking at a DEMOS/OECD conference on 18 May 2004, David Miliband, UK Minister of State for School Standards, outlined his intentions to use 'personalised learning' to raise standards in schools. 'The biggest driver for change', he said, 'is assessment for learning and the use of data and dialogue to diagnose every student's learning needs'. Others, in contrast, have called for the abolition of national tests as unreliable for the task (Wiliam, 2001c). The purpose of this paper is to ask 'how reliable is the data for summative assessment and predicting the future performance of individual children?' in particular in respect of the Key Stage 2 test data.

In England and Wales children are tested at age 7, 11 and 14 on the progress they have made in the National Curriculum at the end of Key Stages 1, 2 and 3 respectively. The results are seen by the Qualifications and Assessment Authority (QCA,

*Corresponding author. Institute of Education, University of Stirling, Stirling, Scotland, FK9 4LA. Email: l.e.doyle@stir.ac.uk

ISSN 1474-8460 (print)/ISSN 1474-8479 (online)/05/010029–17

DOI: 10.1080/14748460500036086

1999), the body which sets, trials and administers the tests, as having a variety of functions. These include the assessment and prediction of the progress of individual children as well as the monitoring of standards at school, local and national level. There is a growing tension between the use of the test results and teacher assessment whereby the QCA (2002) has also adopted the Assessment Reform Group's (2002) advice, which distinguishes between assessment *for* learning via classroom assessment and the assessment *of* learning via grading and reporting. This advice in the form of ten principles, which should guide assessment, has been issued to classroom teachers but the continuing extensive reliance on Key Stage test data for high stakes evaluative purposes has ensured a greater status for them than teacher assessments (Newton, 2003). This is well illustrated by a recent Office for Standards in Education report on assessment in secondary schools (Ofsted, 2003). Although the classroom is identified as 'The home of assessment' (p. 9) in its advice on 'good practice' (p. 6), the Report gives prominence to the Key Stage 2 test data and the examples of good practice it cites are where schools had made extensive and detailed use of the data.

Such close adherence to the advice being given by both the QCA and Ofsted assumes that the test data is reliable. There is evidence to suggest that this may not be the case. Wiliam (2000a, b; 2001a, b, c) challenged the dominant role of the Key Stage test results on a number of grounds. Of particular interest in the context of this paper are his findings that standardised test data, which may be reliable for assessing average abilities for groups of children (for example, sets, classes, schools, cohorts) have questionable reliability for assessing the attainment of individual children and for tracking their progress. Such a contention is especially significant in relation to the Key Stage 2 tests because when the children transfer into Year 7, the secondary teachers are expected, despite their own reservations about the reliability of the tests (Doyle, 2002; Schagen & Kerr, 1999) to use the results as a 'baseline' assessment for measuring the children's progress through Key Stage 3. These findings raise concerns that there may be too much reliance on a form of assessment that still requires further testing. This study is particularly concerned with the variation and unevenness of the results for individual children over time.

Newton (2003), an advocate of national testing, has argued that the reliability of the Key Stage 2 tests is now improved as a result of the QCA's 'test re-test' method but even he agreed with Wiliam that there is the need for more studies of test re-test reliability.

The reliability of both standardised tests in general and the national tests in particular has been questioned by other writers. These have also drawn attention to the need for more work on test reliability and provide further evidence of the possible dangers of relying too heavily on the Key Stage 2 data. Nicholls and Smith (1998) attack standardised tests in general for a lack of transparency and 'argue for a reconceptualisation of reliability that reflects the importance of the theoretical expectations of the test specialist and the learning and solving of the test takers' (p. 34) and that such a requirement should be 'formalised in official standards for test use' (p. 35). Morrison and Wylie (1999) drew a similar conclusion, suggesting that insufficient transparency is displayed and calling for the QCA to adopt a code of standards similar to the American Educational Research Association's (AERA) *Standards for Educational and Psychological Testing* (AERA *et al*., 1999) to ensure a more acceptable level of transparency. Morrison and Wylie's (1999) argument relates to the use of the data for measuring standards in schools. Their concern is that the National Curriculum levels of achievement (for example, at Key Stage 2 pupils are expected to achieve level 4) create the impression that standards in schools can be measured objectively when that level is based on numerical, rather than psychological measurement, but their assertion could also be applied to individual pupils. In their study on the 'high stakes' grammar school Transfer Tests in Northern Ireland, Gardner and Cowan (2000) conclude that 'No attempt is made to make candidates, parents and schools aware of the fact that all Transfer Tests misclassify pupils; that no Transfer Test can measure with greater accuracy than ± 1 grade' (p. 50).

Other writers also specifically refer to the problematic nature of assigning levels or grades to individual children. Black (1998) reiterated Wood's (1991) finding that even with a reliability coefficient that is high enough to be commonly regarded as acceptable (say 0.85 to 0.9), the errors in pupils' scores that are implied may mean that a significant proportion are given the wrong grade. Wiliam's simulation (2001a) suggested that, with a coefficient of 0.95, one might expect between 10% and 24% of students to be differently classified. In a recent paper, Rogosa (2003) effectively illustrates the concept of underestimation of error by showing, in terms of percentile rank, probable true score hit-rate and test re-test results. Yet in his defence of the National Curriculum tests, Newton (2003) finds that small mark differences can certainly result in 'different classification'. However, he rejects the description of 'misclassification' other than in relation to large mark differences. In fact he writes that 'to report only marks … would magnify "misclassification" greatly' (p. 109) because at the point of mark allocation 'misclassification' can also occur as there are more possible marks than there are levels, which consist only of boundaries between sets of marks.

Morrison and Wylie (1999) assert that although teachers and educationalists criticise the tests loosely as 'flawed', 'proponents of the tests cannot rebut such accusations by pointing to comprehensive reliability studies, since none exist' (p. 93). Gardner and Cowan (2000) draw attention to AERA's *Standards* and point out that whilst many countries in the world have testing and assessment regimes governed by them 'the British examination bodies have always avoided the publication of data bearing on the technical fidelity of their assessment instruments' (p. 51). Newton, too, emphasises the need for 'further empirical and conceptual groundwork aimed at reaching a degree of reliability that is acceptable and unacceptable for the uses to which national curriculum tests are put' (p. 93).

What all these critics have in common is that their findings cast doubt over the reliability of standardised tests and their ability to measure accurately the level of achievement of individual children. This scepticism is also to be found at school level. Whilst schools are expected to, and do, use the Key Stage 2 test results for predicting the expected progress of individual pupils, they also continue to demonstrate a lack of confidence in the tests. This is evidenced by the widespread use of Cognitive Abilities Tests (CATs), which schools buy in. Year 7 pupils sit the test to provide an assessment of the pupils' capabilities on intake, with some schools even arranging for the tests to be taken in Year 6 before prospective pupils transfer (Ofsted, 2003, p. 7). The tests provide a set of measures of the individual's ability to use and manipulate abstract and symbolic relationships. They have been shown to be more reliable than Key Stage test data for both predictive purposes at the end of Key Stage 3 and for value added purposes (Moody, 2001), yet in 1997, the Schools Curriculum and Assessment Authority (SCAA), the forerunner of the QCA, had told schools that the CATs were unnecessary because the Key Stage 2 data provided reliable predictors for end of Key Stage 3 achievement. Even in 2002 (Ofsted, 2002) the use of CATs was still considered unnecessary because 'it represents considerable duplication of effort when Year 6 pupils have already been assessed in most aspects of the core subjects at the end of Key Stage 2' (p. 2).

More recent advice, far from insisting that schools abandon CATs, suggested that their use was to be encouraged. This is evidenced by the citing of good practice in one school where CATs were used in assessment procedures. Such advice can also be viewed as tacit acceptance that schools perceive the Key Stage 2 test results as unreliable. Nonetheless, the Key Stage 2 data are still given the greatest prominence with pressure on schools to make even more 'searching' use of them (Ofsted, 2003).

Such increasing emphasis on the use of the Key Stage 2 test results as an assessment tool for measuring the progress of individual pupils seems to be paying insufficient attention to the research as well as the experience of the large number of secondary schools that question the reliability of the tests. This paper focuses on the reliability of the Key Stage 2 tests for assessing and predicting the progress of individual children from the summer term of Year 6 through to the summer term of Year 7.

Methodology

Sampling

The sample consisted of 756 children from nine secondary schools (see Appendix 1). Three groups with between 25 and 32 children per group were selected from each school according to the criteria of sex (except in the single sex schools) and ability. The Key Stage 2 (hereafter referred to as Year 6) test results were used to ensure each group was of mixed ability, although this method did mean that those who had not been entered for the tests because their teachers had assessed them as below level 3, were not included. Each group of Year 7 children was then randomly assigned to take tests in English, mathematics or science in the autumn, spring and summer terms with each group taking the tests for the same subject throughout the testing. Past Key Stage 2 test papers from 1996, 1997 and 1999 respectively were employed in each of the three terms and the results (hereafter referred to as Year 7 tests) were compared with the children's 1998 Year 6 test results.

Table 1. Achieved total sample sizes for each subject in each round of testing

Term	Mathematics	Science	English	Total
Year 6 summer term	220	238	234	692
Year 7 autumn term	236	252	229	717
Year 7 spring term	229	243	227	699
Year 7 summer term	213	225	221	659

Note: n=756

Table 1 (above) shows the achieved sample sizes for each subject in each round of testing. The differences in the numbers from test to test within each subject are accounted for by either the inaccessibility of the Year 6 data, schools' inability or unpreparedness to set the tests in a particular term, or absences from test to test. As a result of the attrition and expansion in the sample, the form of statistical analysis needed to be flexible enough to cope with large amounts of missing data. This was one of the main reasons multilevel analysis was chosen.

Testing

To ensure comparability of marking standards from test to test, the papers were marked by examiners who had all been involved in the training for, and marking of, the tests for the appropriate years for the study (1996, 1997, 1998, 1999).

Another matter concerning comparability was more difficult to resolve. The Year 6 test results for individual children are reported as levels 3–5. For this study, the scores rather than levels were used to compare the Year 6 test results (summer 1998) with the results of the testing in Year 7 (autumn 1998, spring and summer 1999). This is because employing test levels as the only measure of performance would have allowed only the crudest of comparisons over time. The large number of marks within each level results in a big difference in attainment between a child at the top of a level and one at the bottom whereas the scores provided the potential for greater detail. Table 2 (below) illustrates the test levels and raw scores for the four tests used in this study for science. For example, to achieve a level 3 in 1996 children needed to score between 23 and 44 marks, with the result that the same level could be assigned to two children with a difference of 21 marks between the scores.

A further problem with the levels and marks arose because the boundaries are set at different cut-off points each year. Table 2 also illustrates this. For example, whilst for level 3 in 1996 the marks ranged from 23 to 44, in 1998 they ranged from 24 to 41.

For consistency with the type of analysis envisaged, an attempt was made to base the comparison between tests from different years not on the official grade boundaries employed by the QCA but on the statistical equating data employed in the development of the tests. The purpose of this data is to ensure consistency from year to year. Unfortunately, these data were not available despite considerable efforts to

Table 2. Science Key Stage 2 test levels and raw scores for 1996, 1997, 1998 and 1999

	Raw Scores			
Level	1996	1997	1998	1999
2	20–22	18–20	21–23	18–20
3	23–44	21–40	24–41	21–41
4	45–63	41–60	42–61	42–62
5	64+	61+	62+	63+

locate them. Instead, for this study all total scores for the tests, which were applied in the three terms in Year 7, were adjusted for comparability with the 1998 (baseline) scores by equating the level boundaries and using a piecewise linear scaling for intervening scores. So for example, whereas in 1998 the science level 3 score went from 24 to 41 (a difference of 17) and the 1996 level 3 score went from 23 to 44 (a difference of 21) using piecewise linear scaling, a level 3 score of 25 in 1996 (i.e., two marks above the grade minimum) would result in 25.62 across both years (2×17÷21=1.62 +24=25.62). These adjusted results were then analysed using multilevel analysis.

Quite apart from the matter of comparability of scores from test to test, was that of the comparability of the tests themselves. The reliability of the national tests has been criticised and of particular relevance to this study is the suggestion that the tests have been ‘getting easier’ (Hilton, 2001). Such criticisms have undermined both government claims, based on official data, that the Key Stage 2 test results show that standards are rising (Tymms & Fitz-Gibbon, 2001) and also the conclusions of the Rose Inquiry (Rose, 1999) on the 1999 Key Stage 2 tests, set up in response to earlier allegations that the tests were becoming easier. The Rose Inquiry had found that the tests were just as difficult as in the previous year. Renewed suspicions regarding reliability are potentially a problem in the interpretation of results in the present study.

Massey *et al.* (2003), in their report carried out for the QCA concerning the equivalence of standards set in national tests from 1996 to 2001, concluded that there was little evidence to suggest that the 1996 mathematics, and some evidence that the science, Key Stage 2 papers were more difficult than those set in 1999. However, ‘In KS2 English, the experimental evidence indicated that a significant proportion of the apparent improvement in national results may have arisen from variation in test standards indicating that failure to match changes in level thresholds to changes in the relative difficulty of the reading element led to these differences in KS2 English test standards’ (p. 226). It is difficult to estimate the impact of these findings on the study reported here, particularly as the writers were unable to pinpoint which year these changes may have occurred and how the changes were distributed.

Research by Tymms and Fitz-Gibbon (2001) compared reading trends from 1975 to 2000 and found that the Key Stage 2 test results showed a significantly greater increase in standards than other equivalent tests (p. 160). They gave a detailed account, based on Quinlan and Scharaschkin (1999), of the mechanisms used by the QCA to maintain standards over the years. Two of these mechanisms, according to Tymms and Fitz-Gibbon, were cause for concern and they are also relevant to this study.

The first cause for concern relates to the way in which the QCA, whilst they used an anchor test as one of their mechanisms in their attempt to equate standards, restricted its use to just the current and the previous year's tests. Massey *et al.* (2003) found that in Key Stage 2 English and science, though not in mathematics, there were signs that a small part of the very large improvement in national test results reported between 1996 and 2001 may have been a product of a shift in test standards. One of the problems they highlighted was that 'The current year-on-year equivalence is an inherently weak strategy, in which the dangers of incremental shift are readily apparent' (p. 232). Tymms (2004) supports this conclusion but notes that the QCA have changed the way tests are standardised. He refers to the steady rise in test results (with one hiccup) from 1995 to 2000 as Phase 1, and their becoming 'abruptly flat' post 2000 as Phase 2. He concludes that 'the shift from equating standards only to the previous year to maintaining standards over several years happened in 2000–2001 and largely accounts for the different pattern of results in Phases 1 and 2' (p. 491). This could have undermined the use of the tests in the study reported here but much of the researchers' argument concerned grade boundaries and borderline marking of the papers. Here, the concern is not with grade boundaries and levels but with actual scores so in this case the impact of changes from 1996 to 1999, the years from which the test papers in this study were taken, is likely to be less significant.

Tymms and Fitz-Gibbon's (2001) other concern was that of the equating of pre-test, adrenaline-free results to live results. When equivalents are created, based on the results of tests taken under circumstances that are not comparable to the real ones, standards could be lowered (p. 163). As they further point out, this will apply not just to the statistical procedures employed but also to the script examinations. Whilst it should be acknowledged that the Year 7 tests in this study would not have been accompanied by the anxieties of staff and parents that may, during the Year 6 tests, have been conveyed to the children, measures were adopted to counter as far as possible any effects on the children of 'adrenaline-free' testing. The teachers agreed to use the Year 7 test results as part of their own assessment and also to adhere to the QCA expectations regarding invigilation. They further undertook to ensure that the children were made aware that the tests were significant and to aim for a similar gravitas to that of the Year 6 tests the children had taken the previous summer.

Another methodological consideration relating to the use of past Key Stage 2 papers for the testing in this study was that some children might repeat the work in Year 7 that they had covered in Year 6, thereby receiving 'revision' not available to other children. This was likely to vary from school to school as, too, was the possibility that some Year 7 children had already seen the past test papers for revision purposes prior to their Year 6 tests. From the opposite perspective, Stoll *et al.* (2003) suggest that tests, which were designed for the Key Stage 2 curriculum, may not be appropriate to use with the Year 7 curriculum, particularly for low attaining pupils and especially in mathematics. They further note that 'This may be compounded by Year 7 teachers not being familiar with the demands of KS2 tests' (p. 102). Whether children have just studied a topic again as part of the Year 7 curriculum or not studied it since Year 6 will mean that only a proportion of tests set in Year 7 will be of recent learning; the rest will be of retention. However, this consideration may be more appropriate in the context of the suitability of content-oriented Key Stage 2 tests as indicators of performance.

Whilst acknowledging the controversy over the reliability of the Key Stage 2 tests for measuring progress over time, and the other concerns, it was considered that the advantages of the tests outweighed any disadvantages, not least that they are national tests which do carry some authority. The QCA cannot attribute any changes in scores after transfer to discrepancies between the tests because they defend their reliability. However, others may be able to do so.

Multilevel analysis

Multilevel regression was used to model the scores attained by the children in each test. This made it possible to see whether there was any evidence that, for example, girls did better than boys, older children did better than younger, children with higher scores in the 1998 Year 6 tests did better than others. It also made it possible to do this whilst making allowances that the three test results for an individual child can be expected to show some random variation; that different children with the same 1998 Year 6 score can be expected to vary randomly in their later performance and finally that different classes in different schools can be expected to show some random variation. Although the children in this study were grouped in classes, this level was omitted because per subject each school contributed one class only. This meant that, for this study, class and school level were the same.

Multilevel analysis makes it possible not only to test assumptions about the relationship between the response (dependent) variable and the explanatory (independent) variables but to do so at several levels. In this study, the response variable is the test scores. Two types of explanatory variables are employed in multilevel analysis, as explained above. The fixed effects looked for here are: the terms when the tests were set in Year 7, the Year 6 individual scores, the sex and the age of the children, and the number of days' delay in setting the tests (due to the difficulties of ensuring that tests were set on the same day in the different schools). The random effects looked for were on three levels. The first were the effects of the differences between all the test scores for an individual pupil. These were also analysed in the context of the differences between pupils, which were in turn analysed in the context of the differences between schools. Each of the explanatory variables included in the final model was introduced one at a time, and this was done until the model's goodness of fit ceased to change (Plewis, 1997, p. 3).

By taking the various types of random variation into account, the multilevel regression model deployed made it possible to calculate the accuracy of the parameters estimated, more accurately, and thus to ensure that the evidence of the data was not exaggerated. It provided a more accurate insight into the implications of the children's scores than would have been possible with a simple comparison of scores which had not been adjusted to take account of the nested data.

Results of the testing

English

The model for English (see Table 3 below) indicates that an average pupil in an average school with an average summer 1998 Year 6 score of 54.6 (the mean for this sample) would score 50.7 in autumn 1998 (dropping a considerable number of points), 54.7 in spring 1999 (returning to the original score) and 55.0 in summer 1999 (making no further progress). The 'Adjustment for individual summer 1998 score' shows that for each mark above or below the summer 1998 sample mean, Year 7 tests scores were on average increased or reduced by 0.83, indicating that the changes in the pupils' scores were within acceptable limits of random variation for the purposes of measuring average progress across the sample. A number of explanatory variables were explored, such as age and gender, and some of them made technical improvements to the model. However, they made no detectable difference to the key figures and are not discussed here. The standard errors give an indication of how far these figures might be expected to vary if different samples of the same kind from the same population had been taken. Clearly the picture given here is robust.

The variance shown in the model is open to a number of interpretations depending on your view of ability, attainment and testing. Table 3 shows variances and their standard errors, as calculated by the software (MLWin), but also includes standard deviations for ease of comparison. In this case the school level standard error is as large as the estimated parameter, indicating that differences between schools were not only small but also very poorly estimated on the basis of this sample so for the purposes of this study they can be ignored. To understand the pupil level (33.5) and test level (42.5) variance it is easier to consider the standard deviations, produced by taking square roots of the variances: 5.8 at pupil level and 6.5 at test level.

Secondary school teachers are expected to accept the Key Stage 2 Year 6 test mark as accurate and reliable. It is against this mark that any progress produced in the secondary school will be judged. A pupil who scored highly in the Year 6 tests in some sense 'ought' to do well in later tests. From such a perspective the teacher will find that, quite apart from any average drop in attainment between leaving primary school and starting secondary school, pupils with the same Year 6 score will vary apparently randomly in the attainment they show when they arrive in Year 7. The pupil level standard deviation of 5.8 suggests that about a third of pupils will

Table 3. Details of the multilevel regression model for English total scores

Fixed effects	Parameter		(Standard error)
Autumn 1998 adjusted score	50.7		(0.73)
Spring 1999 adjusted score	54.7		(0.76)
Summer 1999 adjusted score	55.0		(0.76)
Adjustment for individual summer 1998 score	0.83		(0.035)
Variance			
		(SD)	
School level	1.7	*1.3*	(1.7)
Pupil level	33.5	*5.8*	(5.0)
Test level	42.5	*6.5*	(3.2)

Note: Year 6 mean test scores = 54.6

diverge by more than 5.8 from what an average pupil with the same Year 6 score would be expected to show. Passing from test to test throughout the year, the teacher will find that pupils, apparently randomly, do better in some tests and worse in others even allowing for the whole cohort making some progress and even by their own individual standards. The test level standard deviation of 6.5 suggests that in any test something like a third of the pupils will diverge by more than 6.5 from what an average pupil with the same level of attainment would be expected to achieve. In other words, the progress an individual child makes from test to test is random.

From the point of view of an outside observer, the Year 6 mark is no different in status from the other test scores. Such an observer might think in terms of each pupil having reached a certain standard, which will carry him or her along with the general progress of the cohort but, depending on the level of that standard, will keep him or her above or below the cohort average by a certain amount. Test scores (including the Year 6 tests) could then be seen as an attempt to measure that attainment by identifying the position of the pupil relative to the rest of the cohort. Any randomness in the results could then be attributed to 'measurement' error on the part of the tests. From this point of view the pupil level standard deviation of 5.8 suggests that for about a third of pupils the Year 6 score estimates their attainment within an error term exceeding ± 5.8 points. The test level standard deviation of 6.5 suggests that each subsequent Year 7 test estimates their attainment within an error term exceeding ± 6.5 for about a third of pupils, indicating that the Year 6 tests were only slightly less unreliable than the Year 7 tests as a measure of a pupil's attainment.

In either interpretation the pupil level standard deviation is a measure of how far it would be unwise for a secondary teacher to take the Year 6 score as a basis for judging a pupil's attainment or ability.

Mathematics

The model (Table 4 below) indicates that an average pupil in an average school with an average summer 1998 Year 6 score of 46.8 (the mean for this sample) would score 43.3 in autumn 1998 (dropping a considerable number of points), 47.0 in spring 1999 (returning to the original score) and 47.0 in summer 1999 (making no further progress). In this there is little difference between mathematics and English.

In the rest of the model, however, the differences from English are vast. For each mark above or below the sample mean in the Year 6 tests of summer 1998, Year 7 individual scores were on average increased or reduced by 0.07—i.e., by a very small amount (compared to 0.83 in English). This suggests that the Year 6 score for mathematics makes virtually no difference to later progress—that is, it cannot be used to predict a pupil's progress. This must be seen against the background of very much larger standard deviations than for English: 7.3 at school level, 12.8 at pupil level and 4.9 at test level, in other words, there was far more deviation from the average score at each level of analysis, particularly at pupil level.

A number of the explanatory variables were explored, resulting in a more complex model, but they made no detectable difference to the key figures in the fixed part of the model. They are also fairly obscure: the pupils' mean score in the summer 1998 Year 6 science, the age of pupils in days, the date on which the spring 1999 test was taken and (in the spring of 1999 only) the sex of pupils. These explanatory variables are unlikely to impact upon the perceptions of a secondary school teacher so they are not further discussed here. Thus the more complex model is omitted in favour of the one above in the interests of clarity.

The school level random variation can be interpreted either as some schools being more successful than others in reducing attainment loss over the interval

Table 4. Details of the multilevel regression model for mathematics total scores

Fixed effects	Parameter		(Standard error)
Autumn 1998 adjusted score	43.3		(2.61)
Spring 1999 adjusted score	47.0		(2.61)
Summer 1999 adjusted score	47.0		(2.61)
Adjustment for individual summer 1998 score	0.07		(0.027)
Variance			
		(SD)	
School level	53.4	*7.3*	(28.3)
Pupil level	164.6	*12.8*	(16.6)
Test level	24.5	*4.9*	(1.8)

Note:Year 6 mean test scores = 46.8

between primary and secondary education or as some schools having received more pupils whose Year 6 test scores were misleadingly high. Both in the model reported above and in the more complex model the pupil level variation (164.6) exceeded the test level variation (24.5), though much more markedly in the model described here. This can be interpreted as much more variation in what happens to pupils' mathematics attainment during the break than English, on the one hand, or on the other as the Year 6 tests being worse, possibly much worse, than tests in Year 7 at measuring a pupil's relative attainment. Either of these interpretations suggests that the preparation for the mathematics Year 6 assessment, more so than for English, produces scores which are virtually worthless as a guide to what a pupil will be capable of in Year 7. This analysis gives some support to teachers who treat all their Year 7 pupils the same without reference to Year 6 achievement. Note that the standard deviation at pupil level (12.8) is far greater than the drop in mean score from 46.8 to 43.3. Quite a large number of pupils actually improve their score over the summer.

Science

The model (Table 5 below) indicates that an average pupil in an average school with an average summer 1998 Year 6 score of 45.9 (the mean for this sample) would score 43.7 in autumn 1998 (dropping a couple of points), 43.6 in spring 1999 (making no further progress) and 47.4 in summer 1999 (improving upon the original score). Although this pattern differs from both that of English and of mathematics, the rest of the model differs very little from that of English and as such also raises the question of the reliability of the Year 6 score as the basis for judging a pupil's attainment or ability.

Table 5 Details of the multilevel regression model for science total scores

Fixed effects	Parameter		(Standard error)
Autumn 1998 adjusted score	43.7		(0.73)
Spring 1999 adjusted score	43.6		(0.76)
Summer 1999 adjusted score	47.4		(0.76)
Adjustment for individual summer 1998 score	0.76		(0.035)
Variance			
		(SD)	
School level	8.6	*2.9*	(4.8)
Pupil level	30.4	*5.5*	(4.1)
Test level	28.4	*5.3*	(2.1)

Note: Year 6 mean test scores = 45.9

Discussion

The present investigation suggests that the Key Stage 2 tests are an unreliable source of information about an individual pupil's progress over time. This is indicated by the randomness of the results. This randomness at an individual level is to be seen as quite distinct from any average pattern for the sample as a whole, for example the 'dip' in attainment for all three subjects when comparing the summer 1998 Year 6 scores of children with the results of the first stage of re-testing in the first term of Year 7 in secondary school. The school transfer 'dip' is the subject of a separate paper.

In English, the randomness indicates that, with a standard deviation of 5.8 at pupil level and 6.5 at test level, the test results are unreliable for around one third of the pupils who took the 1998 Year 6 tests when compared with the results of subsequent tests taken in the autumn, spring and summer terms of Year 7. The difference between these standard deviations suggests that the Year 6 tests were only slightly less inaccurate than the Year 7 tests as a measure of these pupils' attainment. A very similar picture emerged in the science model.

The mathematics model produced a more complex result, which in turn can be subjected to a number of interpretations. The large standard deviation (7.3) at school level could be down to the success of some schools compared to others in reducing the effects of transfer on progress or it might just have been that other schools received more pupils with scores too high to be an accurate measure of the pupils' attainment. The pupil level standard deviation of 12.8 is actually greater than the drop in mean score between the summer 1998 (46.8) and autumn 1998 (43.3) showing that a high proportion of pupils actually improve over the summer. Whether this is seen as a likely indication that pupils' attainment in mathematics over the summer is prone to more variation than in the other subjects or that the mathematics Key Stage 2 tests are even worse at measuring a pupil's relative attainment than those for English and science is open to question.

This paper is concerned with the reliability of test results for individual pupils. Whilst a 0.8 reliability rate which the QCA claims for the Key Stage 2 tests is acceptable at national level, because as Wiliam (2001a) has pointed out, 'the effects of unreliability operate randomly' so that 'averages across *groups* of students … are quite accurate', it is this very randomness that makes for an unacceptable degree of unreliability for individual pupils. In this study, the reliability rate was under 0.7 for English and science and even lower for mathematics—in contrast to Newton's (2003) assertion that the mathematics and science tests were more, rather than less, reliable than those for English at national level.

Wiliam (2001a) also drew attention to a number of compounding factors, which can serve to decrease even this reliability for individual pupils. The usefulness of a test for predicting future performance depends on the correlation between the scores on the test (the predictor) and the scores on whatever it is that is being predicted (the criterion). Generally, in educational testing, a correlation coefficient of 0.7 between predictor and criterion is regarded as good. However, this coefficient can be reported after 'correction for unreliability' based on estimated individual 'true scores'. As such scores are never known—they are statistically calculated—the calculation itself can introduce another element of unreliability because the result will appear to be much better than is possible in practice. Using such scores to place children into sets, as is commonly the practice in classrooms, can produce further inaccuracies for the individual pupil, according to Wiliam, though this study did not investigate the effects of this practice either at individual or school level.

Regarding the use of levels, rather than scores, for the Key Stage test results, the more precision that is exercised in this process, the lower the accuracy is likely to be because at each 'between level' there is room for error and because in any event, the unreliability of the test may have produced an inaccurate 'observed score' as opposed to the unobservable 'true score'. As Wiliam (2001a) stated: 'Firstly, the difference in performance between someone who scored level 2 and someone who scored level 1 might be only a single mark, and secondly, because of the unreliability of the test, the person scoring level 1 might actually have a higher true score' (p. 19). Even if the claimed reliability for the tests is 0.8, increasing the reliability has a very slow impact on the extent of misclassification. Wiliam worked out that 0.60 reliability produces 27% misclassification, whilst even 0.95 still results in a 10% misclassification.

Conclusion

The findings from this study suggest that the Key Stage 2 test results may not be sufficiently reliable for them to be used by secondary teachers to assess and predict the achievement of individual pupils over time, thereby giving support to the teachers' reservations (Doyle, 2002; Schagen & Kerr, 1999). The factors which compound this unreliability, whilst manageable at group level, at the level of the individual could produce a layering effect of unreliability such that the expectations of that individual are seriously either under- or over-assessed. Wiliam (2001a) concluded that the need to monitor standards over time and the need to assess the individual pupil could not be met by national testing because of its unreliability. Instead, he advocated a national anonymous monitoring system using samples of children for measuring national standards over time. For individual children the focus was placed on teacher assessment in the classroom. In England, the Assessment for Learning has placed teacher assessment at the forefront of its advice to teachers but whilst the results of the Key Stage tests are still the focus of public attention, such advice will not be given the status the advisers, from their pedagogical perspective, presume it to have.

This is of particular importance at Key Stage 2 because these test results, for the majority of children, are transferred from primary schools to secondary schools. This makes it difficult for teachers to use other forms of assessment along with the Key Stage results to confer about individual children. This is less of a problem for Key Stages 1 and 2 teachers in primary schools, and Key Stage 2 and 3 teachers in secondary schools. At least potentially, within primary and secondary schools teachers can develop 'communities of practice' and employ a variety of assessment methods in respect of individual achievement and progress. Ruth Sutton (2001) has referred to Key Stage 2 to 3, on the other hand, as 'the muddle in the middle', and Doyle and Herrington (1998) found there was a lack of communication on curriculum and assessment between primary and secondary teachers. Even the 'communities of practice' within schools, however, are difficult for teachers to develop because, as Hall and Harding (2002) found, the pressure to raise standards keeps the emphasis on assessment firmly under the influence of the national tests.

Miliband's (2004) call for 'personalised learning' and his emphasis on the use of assessment data as a means of achieving this, raises concern given the continuing debate over the reliability of the tests in general and the problems associated with Key Stage 2 tests in particular. Whilst the methodological difficulties of administering Year 6 tests in Year 7 cannot be disregarded (Stoll, 2003) this paper nonetheless provides further evidence that more work is needed by the QCA on both the reliability of the tests and the ways in which the results are employed. The threats to reliability have been shown here to be significant and they need to be researched anew when new examination systems and procedures are introduced—before they are introduced, that is, not afterwards. This paper had added to the mounting evidence against the reliability of the National Curriculum tests and the call for the tests to be withdrawn.

Notes on contributors

Lesley Doyle is a Research Fellow in the Institute of Education at Stirling University. Her research interests include transitions across the lifespan, competency-based training and learning regions. She is currently working with the Observatory Pascal and on an evaluation of nurse prescriber training.

Ray Godfrey is a Senior Lecturer in the Department for Educational Research at Canterbury Christ Church University College and works in the Centre for Physical Education research. His work has covered a number of fields, but his particular interest is in removing barriers between qualitative and quantitative analysis.

References

AERA, APA & NCME (1999) *Standards for educational and psychological testing* (Washington, DC, American Psychological Association).

Assessment Reform Group (2002) *Assessment for learning: ten principles* (London, Assessment Reform Group).

Black, P. (1998) *Testing: friend or foe? Theory and practice of assessment and testing* (London, Falmer).

Doyle, L. (2002) Continuity and progression from Key Stage 2 to Key Stage 3. Unpublished PhD thesis, University of Kent, UK.

Doyle, L. & Herrington, N. (1998) Bridging the gap: a case study of curriculum continuity at Key Stage 2/Key Stage 3 transfer, *Management in Education,* 12(6), 11–12.

Gardner, J. & Cowan, P. (2000) *Testing the test: a study of the reliability and validity of the Northern Ireland Transfer Procedure tests in enabling the selection of pupils for grammar school places* (Belfast, Graduate School of Education, The Queen's University of Belfast).

Hall, K. & Harding, A. (2002) Level descriptions and teacher assessment in England: towards a community of assessment practice, *Educational Research,* 44(1), 1–15.

Hilton, M. (2001) Are the Key Stage 2 Reading Tests becoming easier each year?, *Reading, Language and Literacy,* 35(1), 10–22.

Massey, A., Green, S., Dexter, T. & Hamnett, L. (2003) *Comparability of national tests over time: key stage test standards between 1996 and 2001: final report to the QCA of the Comparability Over Time project* (London, QCA).

Miliband, D. (2004) Speech to the DEMOS/OECD conference on 18 May 2004. Available at: www.dfes.gov.uk/speeches/media/documents/PLfinal.doc.

Moody, I. (2001) A case study of the predictive validity and reliability of Key Stage 2 tests results, and teacher assessments, as baseline data for target-setting and value-added at Key Stage 3, *The Curriculum Journal,* 12(1), Spring, 81–101.

Morrison, H.G. & Wylie, E.C. (1999) Why National Curriculum testing is founded on a methodological thought disorder, *Evaluation and Research in Education,* 13(2), 92–105.

Newton, P. (2003) The defensibility of National Curriculum assessment in England, *Research Papers in Education,* 18(2), 101–127.

Nicholls, P.D. & Smith, P.L. (1998) Contextualising the interpretation of reliability data, *Educational Measurement: Issues and Practice,* 17(3), 24–36.

Office for Standards in Education (2003) *Good assessment in secondary schools* (London, Ofsted).

Qualifications and Curriculum Authority (1999) *Key Stage 2 assessment and reporting arrangements* (London, QCA).

Qualifications and Curriculum Authority (2002) Assessment for learning. Available at: www.qca.org.uk/ages3–14/66.html.

Quinlan, M. & Scharaschkin, A (1999) National Curriculum testing: problems and practicalities. The British Educational Research Association Annual Conference, Brighton, September.

Rogosa, D. (2003) *How accurate are the STAR National Percentile Rank scores for individual students? An interpretive guide,* Version 2.0, CAT/6 Survey Stanford University, August 2003. Available at: www-stat.stanford.edu/~rag/accguide/guide03.pdf.

Rose, J. (1999) *Weighing the baby: report of the Independent Scrutiny Panel on the 1999 Key Stage 2 National Curriculum tests in English and Mathematics* (London, Department for Education & Employment).

Schagen, S. & Kerr, D. (1999) *Bridging the gap? The National Curriculum and progression from primary to secondary school* (Slough, NFER).

Stoll, L., Stobart, G., Martin, S., Freeman, S., Freedman, E., Sammons, P. & Smees, R. (2003) *Preparing for change: evaluation of the implementation of the Key Stage 3 strategy pilot* (London, DfES).

Sutton, R. (2001) *Primary to secondary: overcoming the muddle in the middle* (Salford, Ruth Sutton Publications).

Tymms, P. (2004) Are standards rising in English primary schools?, *British Educational Research Journal,* 30(4), 477–494.

Tymms, P. & Fitz-Gibbon, C. (2001) Standards, achievements and educational performance: a cause for celebration?, in: R. Phillips & J. Furlong (Eds) *Education, reform and the State: twenty-five years of politics* (London, RoutledgeFalmer).

Wiliam, D. (2000a) Integrating formative and summative assessments functions of assessment. The European Association for Educational Assessment Conference, Prague, Czech Republic, November.

Wiliam, D. (2000b) The meanings and consequences of educational assessment, *Critical Quarterly,* 42(1), 105–127.

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

Wiliam, D. (2001b) *Level Best? Levels of attainment in National Curriculum assessment* (London, ATL).

Wiliam, D. (2001c) What is wrong with our educational assessments and what can be done about it?, *Education Review,* 15(1), 57–62.

Wood, R. (1991) *Assessment and testing: a survey of research* (Cambridge, MA, Cambridge University Press).

Appendix 1. Cluster and secondary school sample

Cluster/LEA type	School	Type of school/ funding	Attainment according to national average	Free School Meals %	Sex	Ethnic mix	Form entry size
A/Selective suburban	1	Secondary modern/ LEA	Below	25	Mixed	Low	6
	2	Grant maintained/ DfEE	Above	Low	Mixed	Low	6
B/Non-selective inner city	3	Comprehen sive/Church of England/ LEA	Below	Above average	Mixed	Very high	6
	4	Comprehen sive/LEA	Below	65	Mixed	Very high	6
C/Non-selective inner city	5	Comprehen sive/LEA	Average	69	Girls	High	6
	6	Comprehen sive/LEA	Below	80	Mixed	Very high	6
D/Selective small town	7	Grant maintained/ DfEE	Below	24	Mixed	White	6
	8	Secondary modern/	Very below	50+	Mixed	White	6
	9	LEA Grammar/ LEA	Above	5	Boys	White	5