British Educational Research Journal
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713406264

Are standards rising in English primary schools?

Peter Tymms [a]
[a] University of Durham, UK

Online Publication Date: 01 January 2004
To cite this Article: Tymms, Peter (2004) 'Are standards rising in English primary schools?', British Educational Research Journal, 30:4, 477 - 494
To link to this article: DOI: 10.1080/0141192042000237194
URL: http://dx.doi.org/10.1080/0141192042000237194

British Educational Research Journal
Vol. 30, No. 4, August 2004

Carfax Publishing
Taylor & Francis Group

Are standards rising in English primary schools?

Peter Tymms*
University of Durham, UK

(Received 9 January 2004; conditionally accepted 10 March 2004; accepted 17 March 2004)

The officially reported impressive rises in standards in mathematics and English in primary schools since 1995 are challenged. The article looks at the increases from four different perspectives. First, the general pattern of change is considered. Secondly, the statutory test data are compared with the results from several different studies. These indicate a complex pattern with clear rises in standards, but not as strong as the official data suggest up to 2000. Since 2000 the official data have shown little change and this is largely confirmed by independent tests. Thirdly, the standard setting procedures are considered and mechanisms by which the observed patterns could have been produced are set out. Finally, parallels are drawn with the experience in Texas where an apparently remarkable set of data was shown to be largely illusory. A case is made for an independent body to be set up with the express and sole purpose of monitoring standards over time.

Introduction

Statutory test data have now been publicly available for the end of primary school education in England since 1995. These data have shown a remarkable rise in standards and have been used to promote a positive perception of the efficacy of government policies both within England and abroad. But for some time there has been concern that the conclusions drawn from a simple examination of the statutory test data are unsafe. A number of different perspectives on standards in primary schools within England have been generated and this article brings that information together.

Michael Barber, Head of the Prime Minister's Delivery Unit, has been particularly active in publicizing the gains. He has done this in an academic journal (Barber, 2001), in professional publications (USA and Internet) as well as in talks and papers. For example, in an American professional publication he wrote, 'Large scale reform is

*Curriculum, Evaluation and Management Centre, Mountjoy Research Centre 4, University of Durham, Stockton Road, Durham DH1 3UZ, UK. Email: P.B.Tymms@dur.ac.uk

ISSN 0141–1926 (print)/ISSN 1469–3518 (online)/04/040477-18

DOI: 10.1080/0141192042000237194

not only possible but can be achieved quickly' (*Education Week*, 15 November 2000), citing the statutory test results as evidence. Others have also accepted the results uncritically, as, for example, in the National Audit Office (2001) report, which takes as one of its examples the introduction of the National Literacy Strategy and uses as evidence for its success the statutory test data (pp. 77–82). The report then proceeded to set out lessons that can be learnt for policy-making more generally.

Clearly, the rise in statutory test scores at the end of primary school in England since 1995 is of considerable importance. It is being used to justify policies and to promote certain ways of working. It is being trumpeted on the international stage and is having a major impact on educational policy generally.

Scope of this article

This article looks at the adequacy of the statutory test data in its role of monitoring standards over time and it does this in four ways: by analysing the pattern of changes over time as reported by the official statistics, by comparing the official results with independent data, by considering the mechanisms for monitoring standards operated by the Qualifications and Curriculum Authority (QCA) and by comparing the experience of England with Texas.

The two areas considered here are English and mathematics. Both sets of tests aim to reflect children's basic skills as taught within the National Curriculum. The English test results deal with reading, writing and spelling. The latter is weighted to contribute 7% to the final English mark and it should be noted that the majority of independent data reported here relate to reading. Whilst writing is clearly an important skill, it is often difficult to assess and no independent assessments have been found relating to children's writing standards at the end of primary school over the last few years. Some analysis was reported in Massey *et al.* (2003) and that is discussed later together with sampling issues and the alignment of the independent tests with national assessments.

Ideal data for monitoring standards

All monitoring systems have imperfections but it may be useful at this early stage to set out the characteristics of an ideal structure. A perfect system for monitoring standards over time would involve the same secret tests used repeatedly on equivalent samples of pupils of the same age at the same time of year. Testing samples rather than full populations makes the process efficient and, compared with national testing, a much smaller operation. A secret test is needed because once the content of a test becomes known by teachers it is hard for them, even if they are so motivated, to keep the ideas in the assessment hidden. Further, if there is any pressure on the schools, there will be enormous temptation to include at least a little of the information related to the test in teaching and/or to prepare the children in some other way. If a different test is used on each occasion then it is necessary to use some statistical procedures to make the tests equivalent. This is hard, hence the well-known phrase 'if you want to measure change don't change the measure'. Finally, any body responsible for tracking standards

should be briefed to create a monitoring system designed to be resistant to the inevitable shifts in curricula and language over time.

The Assessment of Performance Unit (APU), which was established by the Department of Education and Science, carried out much well-respected work along the lines of those described above from the 1970s until it was disbanded in 1990. It might provide a model for a new body, although in the present questioning climate it would make sense for any new unit to be independent and to be guaranteed finance for a sufficient period for it to be able to plan well into the future.

The available data sources

Eleven separate sources of information are used in this article.

1. Statutory end of Key Stage 2 test data (Year 6: 11-year-olds)

These are the official statistics compiled by the Department for Education and Skills (DfES) analytical services and published on the Web (http://www.dfes.gov.uk/performancetables). They are presented each year as the percentage of children who have reached each of the possible levels. The key statistic is the percentage of children reaching at least a so-called level 4. Data are reported for English, mathematics and science. The test material is the responsibility of the QCA, which was known as the School Curriculum and Assessment Authority (SCAA) until 1997. Two sets of new tests are produced every year with one being kept in reserve in case of unexpected problems. Cut-scores are set to identify which pupil has attained which level after the national data become available. The tests are the subject of stringent security and checks are made to ensure that proper procedures are being followed in schools. The tests are marked externally and scripts are returned to schools after marking. Schools can then challenge the grades of individuals.

The DfES publishes the results from the tests as school performance tables. The data show the percentage of pupils in each school attaining a level 4 or above. The Office for Standards in Education (Ofsted) also uses the results in its inspection of schools.

2. Statutory end of Key Stage 3 test data (Year 9, 13-year-olds)

As above, except that these data are not published on a school-by-school basis.

3. PIPS project

Schools and local education authorities (LEAs) can opt to join the Performance Indicators in Primary Schools (PIPS) project run by the Curriculum Evaluation and Management (CEM) Centre (see, for example, Fitz-Gibbon, 1996; Tymms & Coe, 2003). The PIPS project aims to provide feedback to schools for self-evaluation (see, for example, Tymms, 1999; Tymms & Albone, 2002). Very broadly based data are

collected for all ages in primary schools in England and this includes attainment and attitudes in reading, mathematics and science in Year 6 as well as developed ability. In this article the reading and mathematics attainment scores from the same 122 schools and over 5000 pupils each year are reported for the years 1997 to 2002. The tests were specifically written for the PIPS project and are linked to the National Curriculum. In any one year they are found to correlate very well with the Year 6 statutory results (for mathematics in 2003 r = 0.85 and for PIPS reading with English statutory test results r = 0.83).

Similar data are also reported for Year 4 from 1999 to 2003 using the same 382 schools and more than 10,000 pupils per year.

4. The MIDYIS project

The MIDYIS project is another monitoring project run by the CEM Centre. At its heart is an innovative test of developed abilities (http://www.midyisproject.org), which includes a mathematics subtest within it. The MIDYIS test comes in a number of forms, but crucially, much data are collected when pupils start school in Year 7, just one summer holiday after they have taken the statutory tests in Year 6 in primary school. In 2002 the project posted the standardized scores for the mathematics subtest from more than 31,000 pupils each year from 1999 to 2002 on the Web. This test has been shown to have good predictive validity for the Key Stage 3 statutory tests and it is these results that are included in the next section.

5. Davies and Brember

Julie Davies has collected data using the same tests of attainment and self-esteem from the same five randomly chosen primary schools in one LEA since the Education Reform Act of 1989. In Year 6 they used standardized tests of reading and mathematics, chosen with care to represent the skills that they identified as being important at the end of primary education. The results have been presented in a series of articles (Davies & Brember, 1997, 1999, 2001). The latest of these includes data up to 1998 and it is these results that are included in this article.

6. Leverhulme study five-year longitudinal study

Brown *et al.* (2003) studied two cohorts of children as they progressed through primary school. Cohort 1 was tracked from Year 1 to Year 4 whereas Cohort 2 was tracked from Year 4 onwards. This design meant that data could be brought to bear on Year 4 in 1997/98, the start of the Cohort 2 data, and 2001/02, the end of Cohort 1 data. Data were available from 35 schools and about 1300 children at the start and end for Year 4. The sample was broadly based across four LEAs and it concentrated on numeracy related to the National Strategy rather than the National Curriculum. Individual item characteristics were followed across the years and estimates made of the changes in facilities between the two time points for Year 4. Systematic differences

were noted between different areas of numeracy, with some items becoming easier and some more difficult. Average effect sizes were reported for the start and the end of the year. These were almost identical and the mean of the two is used in this article.

7. TIMSS & TIMSS-R

The Trends in International Mathematics and Science Study (TIMSS), formerly known as the Third International Mathematics and Science Study, is coordinated by the International Association for Evaluation of Educational Achievement (IEA) and has sought to generate comparative data on achievement and learning contexts internationally (www.iea.nl). In 1994/95 mathematics test data were collected from representative samples of 41 countries for pupils aged 8/9 (Year 4 in England) and this was repeated with the same cohort, although not the same individuals, in 38 countries for pupils aged 13/14 (Year 9 in England). Twenty-nine countries participated in both studies and Ruddock (2000) notes that the overall performance of the English sample 'did not change significantly from 1995 to 1999'.

8. DfES/QCA/Ofsted

Test results were collected for literacy and numeracy in the summers of years 1999–2001 inclusively for Years 3, 4 and 5 (Minnis & Higgs, 2001). This meant that the data could be looked at to see if there had been a change in the mean test scores over the three years. The age-standardized scores of pupils who moved up through the years could also be compared.

The project involved the same schools over the three years and they were chosen to give representative samples. Two sets of 300 schools were employed but for literacy there was a drop in the number of very low scoring schools participating in the study to the extent that about a third of the lowest category did not participate in 2001. (This was estimated from Table 1.2.1.)

As an aside it is worth noting that the standardized scores for each of Years 3, 4 and 5 for reading, spelling and numeracy were all significantly below 100 in 1999, the first year of data collection. The original standardization procedure would have established a figure of 100 as the average score using a representative sample. The figures in 1999 therefore suggest a drop in standards. But despite many attempts the author has been unable to find when the tests were originally standardized.

The data were analysed in a number of ways, including using multilevel models to view the results over the years. The models are very complex and produce some odd findings, but that is not the concern of this article. There are two relevant sets of findings. The first is that the rise in standardized scores from Year 3 to Year 5 was on average 0.075 SD units in reading per year. The parallel figure for numeracy was 0.09. These are very small changes.

The second concerns the Year 5 reading and numeracy year-on-year results that are used below in the knowledge that low scoring schools for reading were leaving the project over the period of interest. The results are included later in this article.

9. Hilton's textual analysis of the statutory tests

Hilton (2001) looked carefully at the English statutory tests for each year from 1998 to 2000. She found that 'the reading tests were progressively easier for the children to answer … because the number of higher-order reading skills … has decreased each year'.

10. National Foundation for Educational Research

Whilst no year-on-year analysis of the test data has been published, Whetton (*Times Educational Supplement*, 10 May 2002) is quoted as saying in relation to NFER tests: 'If there was a lot of evidence that there was a jump or drop in performance then we would have had to re-standardise. But this has not happened in the past four years'. Without quantification, it is a little difficult to know what exactly is meant by this quotation other than that Whetton's opinion appears to be that the academic standards of children in primary schools has not changed.

11. QCA commissioned comparability study

In 1999 the QCA commissioned the Research and Evaluation Division of the University of Cambridge Local Examinations Syndicate to study the equivalence of standards in the statutory tests over several years. The work was completed towards the end of 2001 and was published in November 2003 (Massey *et al.*, 2003).

The study covers statutory test results at the ends of the first three key stages but only those relating to English and mathematics at the end of Key Stage 2 will be considered here. Data were collected in three separate ways. First, statutory tests from 1996 and 1999 were administered to equivalent groups in Northern Ireland. This was extended to study the 1996 and 2000 tests for English but not mathematics. Secondly, qualitative data were collected from teachers and pupils concerning the tests. Finally, data were collected from LEAs in England that had independent test data for the period in question.

The study suggests that the 1999 English tests were more leniently graded than the 1996 test by 0.28 levels or an effect size of 0.34. A similar but slightly larger discrepancy was found for the English results in 2000 compared with 1996, although the discrepancy between the two sets of results may be explained by sampling variation. The team was able to look carefully at the breakdown of the English levels into the constituent parts over the years and concluded, 'some of the recent improvement in Reading results … are illusory … Conversely standards for marking writing seem to have been maintained' (p. 63). Contrasting results were found for mathematics where they concluded that there 'is no suggestion here that standards … might vary' (p. 71).

The judgements of a 'small' group of experienced teachers were used to look at the quality of scripts from the Northern Ireland work, from pupils who had attained a level 4 in the 1996 or 1999 papers. Although the authors caution care in placing too

much reliance on the exercise, they concluded that the weight of opinion was that the work from 1999 was 'of a lower quality' than that from 1996.

The interviews with children indicated that they 'clearly perceived the 1999 paper to be more accessible and user-friendly than the 1996 version' (p. 147). This has no direct implications for standard setting, although the writers raise the question as to whether cut-scores should be set with such information in mind or not.

Six LEAs provided data on standardized tests that could be linked to KS results. Four of them had reading test data and the researchers concluded that the linked data indicated that 'children with equivalent reading scale scores have obtained better and better statutory test levels … with an uplift of about a tenth of a level per year' (1996–2000) (p. 197).

Two LEAs had independent standardized mathematics scores and a third had quantitative aptitude test data. There appeared to be some variation in the link between these scores and the statutory test results over the years but the authors concluded that the setting standards at the end of the period being studied (2000) were the same as at the start (1996) (p. 212).

The pattern of Key Stage 2 results

The KS 2 results present an interesting pattern that is best viewed graphically in Figure 1.

The results for English show a steady rise from 1995 to 2000 and then the graph flattens. The mathematics data show a very similar pattern, with a noticeable disjunction in 1998 that, at the time, was attributed to the introduction of the oral test. Be that as it may, the line for mathematics resumes its shadowing of the English line the following year.

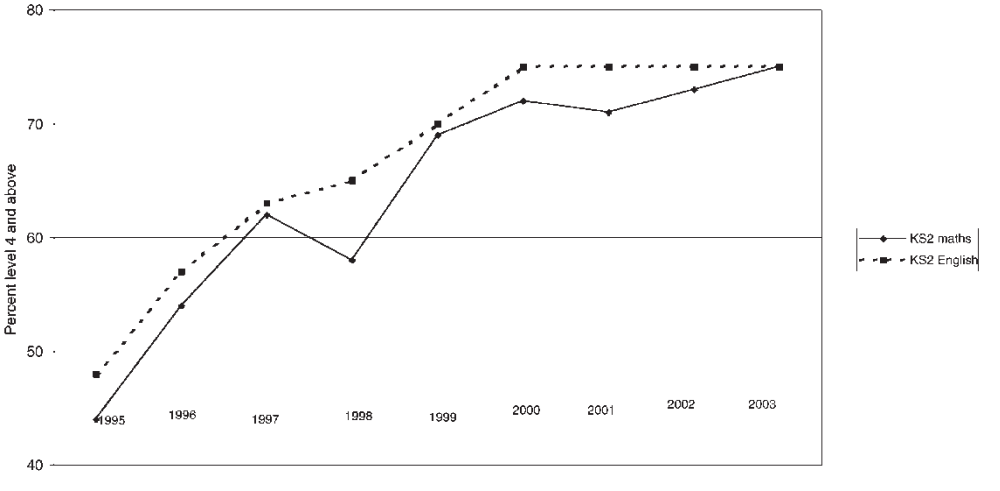

Figure 1. Statutory test results over the years

Whilst each point on the graph is the result of testing almost the full population of state school pupils in England, the error surrounding each point is much greater than would be obtained if each point were a simple average mark. This is because each point is based on a new cut-score, which must be made each year, and this cut-score must be an integer value corresponding to a particular mark. A change of one mark in 1996 would have made about 1.4% difference to the proportion of students being awarded a level 4 or above. In 2000 the figure is 1.8. This can be estimated from the graphs given by Massey *et al.* (2003). In other words, the points on the graph cannot be more precise than 1–2% and it would be quite understandable if the errors were double or even triple those figures. For this reason the apparent slight drop for mathematics in 2001 provides little evidence for any change in standards. Further, the slight departures from the overall patterns can be viewed as being due to errors of measurement.

The general pattern is very clear. For both mathematics and English the results rose dramatically from 1995 to 2000 and then remained steady. The period from 1995 to 2000 will be referred to as Phase 1 and from 2000 onwards as Phase 2.

The patterns in the data are surprising in two ways. Firstly, the discontinuity between Phases 1 and 2 is unexpected. Something quite abrupt must surely have happened. Secondly, it is surprising to see the mathematics and English lines running so parallel to one another. Surely one would be expected to rise more quickly than the other or one to flatten off earlier or for some other patterns to appear. It seems strange that after seven years of change they still show the same relationship to one another. We know from many years of analysis of school-based data that the school influence on some parts of the curriculum is greater than others. Typically schools account for a greater proportion of the variance in relation to mathematics results than for English results. Bryk and Raudenbush (1989), quoted in Teddlie and Reynolds (2000, p. 85), for example, comment in relation to their analysis of data at five different time points from 68 elementary schools, 'Over 80 per cent of the variance in mathematics learning is between schools! These results constitute powerful evidence of school effects that have gone undetected in past research. *As we would expect, the between-school variance in reading is somewhat less*, 43.8 percent, although still substantial' (p. 732, emphasis added). And the effect sizes for interventions from experimental data are typically higher for mathematics than English (see, for example, Fitz-Gibbon, 1992). This suggests that something other than a change in standards is being reflected in some aspects of the data.

Bringing the results together

The data from several different sources are recorded below in Tables 1 and 2. The data were then adjusted to show the changes over the two periods per year in the different data sets (Tables 3 and 4).

Table 1. English and reading test data

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
KS2 English %	*48*	*57*	*63*	*65*	*70*	*75*	*75*	*75*	*75*		
KS2	100	103.4	105.7	106.5	108.6	110	110	110	110		
KS3 English %	*65*	*64*	*63*	*65*	*66*						
KS3	100	99.6	99.2	100	100.4						
Massey *et al.*		100			104.1						
Massey *et al.*		100				103					
D & B	96.8	98.4	97.7	99.6							
PIPS A			100	101							
PIPS B					100	101	101	101			
PIPS Y4							100	100.4	100.9	101.4	101.2
QCA/DfES/ NFER Y5				98.9	100.1	100					
Massey *et al.* LEA 1		98.5	100.4	100.6		101					
Massey *et al.* LEA 2				104.0	104.9						
Massey *et al.* LEA 3			97.9	98							
Massey *et al.* LEA 5				105.9	105.9	106.3					

Notes on Table 1:
(a) The KS2 English results are initially recorded as the percentage of pupils gaining a level 4 or higher and these are then converted to standardized scores with a mean of 100 in 1995 and an SD of 15. (This was done by using tables to find the z score shift needed to create the percentages shown and then converting the z score.)
(b) The D & B (Davies & Brember) data are based on a reading test, which was standardized to have a mean of 100 in 1986.
(c) The PIPS data were standardized to have a mean of 100 in 1997 for form A and 1999 for form B.
(d) The PIPS Year 4 data were standardised at 100 in 1999.
(e) The QCA/DfES/Ofsted study was based on a test, which was standardized at an unrecorded date.
(f) The KS3 results are for the cohorts of pupils who took the KS2 tests in the year stated for the column and standardized to the first year in the table.
(g) Massey *et al.*'s data were used to estimate the true rise by comparing standard deviation changes found in the statutory test results with the Northern Ireland data.
(h) The LEA data are the standardized scores reported in Massey *et al.*

Two issues: sampling and test alignment

In comparing test results with the end of key stage results two issues must be considered. One relates to the representativeness of the samples and the second to the relevance of the tests to standards at the end of KS 2. In view of the very large sample sizes in Tables 3 and 4, errors of measurement are not an important factor.

Table 2. Mathematics test data

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
KS2	*44*	*54*	*62*	*58*	*69*	*72*	*71*	*73*	*75*		
Mathematics %											
KS2	100	103.8	106.8	105.3	109.7	111	110.6	111.5	112		
KS3	*60.1*	*62*	*65*	*66*	*67*						
Mathematics %											
KS3	100	100.7	101.9	102.3	102.8						
Massey *et al.*		100			105.3						
D & B	99.4	99.6	102.6	102.5							
PIPS			100	106	107	109	109	109			
PIPS Y4							100	101.9	102.3	102.2	101.6
MIDYIS					100	101	103	104			
Brown *et al.*				100				102.7			
QCA/DfES/ Ofsted				99	102.2	104.1					
Massey *et al.* LEA 1			101.1	102.1		104.9					
Massey *et al.* LEA 5				106.9	106.9	108.4					
Massey *et al.* LEA 6		99.9	101.1	101.4							

Notes on Table 2: The comments below Table 1 apply to Table 2 mutatis mutandis.

Table 3. First period: changes for reading and mathematics per year

Study	Years	Number of years	Pupils total	English/ reading change	Mathematics change
KS2	95–00	5		2.0	2.0
KS3 matched cohort	95–99	4		0.1	1.7
Massey *et al.*	96–99	3	800	1.4	1.7
Massey *et al.*	96–00	4	1,000	0.7	
D & B	95–98	3	800	0.9	1.0
PIPS A	97–98	1	10,000	1.0	
PIPS B	99–00	1	10,000	1.0	
PIPS	97–00	3	15,000		3.0
MIDYIS (quant. ability)	99–00	1	62,000		1.0
QCA/DfES/Ofsted	98–00	2		0.5	2.5
LEA 1	96–00	4	270,000	0.6	1.3[1]
LEA 2	98–99	1	156,000	0.9	
LEA 3	97–98	1	5,000	1.1	
LEA 5	98–00	2	18,000	0.7	0.7
LEA 6 (quant. ability)	96–98	2	18,000		0.7
Mean of D&B to LEA6 weighted by number of years			0.77	1.54	

Based on 97–2000 only.

Table 4. Second period: changes for reading and mathematics per year

Study	Years	Number of years	Pupils total	English reading	Mathematic change
KS2	00–03	3		0.0	0.3
PIPS B	00–02	2	10,000	0.0	
PIPS	00–02	2	10,000		0.0
PIPS Y4	01–03	2	20,000	0.4	1.1
MIDYIS (quant. ability)	00–02	2	63,000		1.5
Mean of PIPS B to MIDYIS weighted by number of years			0.20	0.87	

NB1. The data from Brown *et al.* do not appear in the table because they cross the first and second periods (98–2002). Overall they showed a rise in mathematics of 0.7 per year.
NB2. The TIMSS data do not appear in the table because they referred to scores for Year 4 and then Year 9 pupils. But it should be noted that the study recorded no change in standards.
NB3. The QCA/DfES/Ofsted changes in pupil scores as they aged also do not appear in the tables, but in standardized scores they recorded rises per year of 1.1 for reading and 1.4 for mathematics.

Massey *et al.*'s data present a special case because randomization was used to create equal groups. Further, the equality of the groups was checked using data from Northern Ireland. This check suggested that the 1996–2000 English and the 1996–99 mathematics results in Table 3 underestimate the true rise.

Davies and Brember's data involved choosing schools at random from one LEA and are therefore broadly based although not necessarily representative of England. The PIPS data have been checked for their representativeness of England, as have the MIDYIS data. The QCA/DfES/Ofsted data were chosen to be representative and Massey *et al.* closely examined the LEA data against the KS2 results. They concluded that reading results for LEA 1 'may not be untypical' (p. 197). The LEA mathematics data were more complicated but presumably LEA 1 may again be taken as representative. The other LEA data are important records of changes for whole LEAs, but as with Davies and Brember's data, may not be representative of England as a whole. This issue is taken up again when discussing the scores generally.

Turning now to the alignment between the tests and the end of key stage assessments, it should be stated from the outset that none was perfect. Even the key stage tests themselves have changed noticeably over the period in question. For example, the study of children's views by Massey *et al.* concluded, 'the children clearly perceived the 1999 paper to be more accessible and user-friendly than the 1996 version' (p. 147). Further, Massey *et al.*'s data which involved repeats of the English statutory test data could be criticized because they were given in a system that had not followed the English curriculum. This was, however, carefully examined and not found to present a problem. More of an issue might be found with the use of reading test data in Tables 1–4 because the English results include writing and to a lesser

extent, spelling. There are, however, at least three reasons to suggest that their study is important.

1. The scores on reading tests are valuable in themselves. They have been widely used for many years and if standards have risen in any meaningful way it would be a strange thing if this were not reflected in reading scores.
2. The reading tests all correlated well with the statutory data. The correlations for LEAs 1, 2, 3 and 5 were 0.72, 0.70, 0.72 and 0.71 respectively. And, as has already been noted, the figure for PIPS was 0.83.
3. Massey *et al.* were able to link standardized scores to the writing and reading components in LEA 5 and wrote, 'any shifts in KS2 English test standards arose largely or wholly in the reading component' (p. 211). Further, on the basis of the data that they collected from Northern Ireland, they concluded, 'some of the recent improvements in reading … are illusory' whereas 'standards … for writing have been maintained' (p. 63). They saw this as contrary to the 'widespread impression'. Some independent confirmation of this finding in relation to reading would be valuable.

The mathematics test data might be expected to be more closely aligned to the statutory test results, and indeed the correlations for LEAs 1, 5 and 6 were 0.78, 0.79 and 0.74. The correlation with PIPS test results was 0.85. The degree of correlation reflects the alignment of the tests with the statutory data and the figures suggest a close correspondence.

Patterns in the data

The general patterns arising from a comparison of the statutory test data and data from 10 independent projects involving nearly half a million pupils were as follows.
(1) Phase 1:

(a) The English and mathematics statutory test data rose by about 2 standardised points (using a standard deviation of 15) a year.
(b) For reading, the independently collected data all showed an increase and this amounted to 0.77 points per year on average. The studies as a whole present a consistent pattern, with only a little variation around this average figure. This lack of variation suggests that even though the samples were not always known to be representative this was not a key factor.
(c) The average reading rise was lower than the first rise in Massey *et al.*'s data and this may be because the latter included writing. As noted, the second figure in Massey *et al.*'s data is thought to be an underestimate.
(d) In mathematics, the independently collected data showed an average rise of 1.54. There was more variation around this figure than was seen for reading. This may be because mathematics tests may be more sensitive to content than reading tests.
(e) The average mathematics rise (1.54) was very close to Massey *et al.*'s figure, which

is claimed to be an underestimate. On the other hand, the TIMSS data indicated no rise at all.

(2) Phase 2:

(a) The English and mathematics statutory test data remained almost constant although there was a slight rise for mathematics.
(b) The one independent source of Year 6 reading data showed no rise, although the pupils appeared to be up by 0.4 points in Year 4.
(c) As for Phase 1, the independent mathematics data show a large variation. The average was 0.87, suggesting that the statutory data may have underestimated the true rise, but this is only a suggestion. Brown *et al.*'s data crossed the two phases but only rose by 0.7 points per year. However, it is difficult to know when the rise occurred.

The mechanism(s) used by QCA to maintain standards

Tymms and Fitz-Gibbon (2001) gave a detailed analysis of the mechanisms used to maintain standards over the years. The analysis was based on a paper by Quinlan and Scharaschkin (1999) which indicated that a number of sources of information were used to set cut-scores each year. These were: marker opinion, professional scrutiny of the test papers (Angoff technique), earlier use of the live test and the employment of an anchor test. Two major difficulties were identified. One has already been noted and that is that the cut-score must correspond to a mark and this limits the potential accuracy of any noted change in standards from one year to the next. It also, incidentally, has implications for the potential of the system to look at changes across the full range of attainment since cut-scores restrict further analysis. The second major difficulty is that attempts were only made to equate standards from one year to the next. Even the anchor test was restricted in its use to this purpose. Quinlan and Scharaschkin note (p. 11) that the test development agencies base their draft level thresholds on four indicators, one of which is 'equating this year's and last year's test via an anchor test'. This leaves the door open to drift over the years since the standard set in any one year can only be of limited accuracy and the next year's cut-score builds on it. One only has to consider the pressures in the system to see how the rise in Phase 1 could be the result of the system rather than any change in standards. This analysis was presented at the QCA on 16 November 2000. Massey *et al.* (2003, p.232) also identify this issue as problematic: 'The current focus on year-on-year equivalence is an inherently weak strategy, in which the dangers of incremental drift are readily apparent'.

In Phase 2 it seems that the QCA tightened its procedures. This is clear in a record of a meeting between the Statistics Commission, the DfES and QCA (personal communication from Gill Easterbrook, Chief Executive of the Statistics Commission, to the author, 17 April 2002) and in an accompanying diagram of the 12 stages of the test development cycle. This clearly refers to checking standards by referring to

information from previous years (plural) and in stage 4 this involves anchor test data. This is quite different from Quinlan and Scharaschkin's Figure 1 (the test cycle), which refers specifically to equating 'to previous year's test'. In other words, care is now taken to set cut-scores on the basis of data collected not just in the previous year but over several years. This immediately deals with one fundamental problem and may well be the reason for the abrupt change between Phases 1 and 2. It still leaves the other difficulties associated with the use of cut-scores to which can now be added another problem. The average percentage reported achieving at least a level 4 has levelled off at above 70%. This means that our nation's system for monitoring standards focuses on the top third to a quarter of the population. The original system focus was close to the average.

The Texas experience

In 1990 the Texas Assessment of Academic Skills (TAAS) started to monitor the state mandated curriculum. The test was very high stakes in that students had to pass it in order to graduate from high school and the scores were linked to important evaluations of teachers and principals. As in England, a new set of tests was produced every year and, as in England, teachers administered the tests, which were marked externally, and the tests were released after grading every year, enabling schools to use past papers to coach their students.

The scores on the TAAS rose dramatically and the trend was dubbed the 'Texas miracle'. The increase was seen both in reading and mathematics and was accompanied by a decrease in the gap between whites and students of colour. As in England, politicians trumpeted the apparent success.

The TAAS is restricted to the state of Texas but there is a federally mandated monitoring system known as the National Assessment of Educational Performance (NAEP). This system is well regarded for its technical quality using well-designed sampling frames, broadly based items in a variety of formats and trained consultants to administer the tests. Klein *et al.* (2000) have used the NAEP data to challenge the validity of the claims surrounding TAAS. They focused their attention on the Grade 4 and 8 results in 1994 and 1998. They concluded, 'over a four-year period, the average test score gains on the NAEP in Texas exceeded those of the nation in only one of the three comparisons, namely: fourth grade mathematics'. With respect to the ethnic claims, they note, 'whereas the gap on the NAEP was large to begin with and got slightly wider over time, the gap on TAAS started off somewhat smaller than it was on NAEP and then got substantially smaller'. In other words, the Texas miracle was shown to be an illusion, although there were some gains in mathematics.

Klein *et al.* suggest that the reason for the illusion has to do with coaching and test preparation. They may be right but one would also want to be assured that the standard setting procedures were very robust. Further, it is difficult to answer the questions: to what extent are the rises the result of teaching test technique, and to what extent are rises due to teaching to the test? One can gain some purchase on the first question by looking at previous work, which indicates 'gaining familiarity with

taking tests results in higher scores, usually of some 3 to 6 standardised points' (Jensen, 1980). This suggests that teaching test technique will have a limited short-term impact on year-on-year test results as teachers train their children to take the tests. To what extent each of the two other factors, the standard setting procedures and teaching to the test, account for the remainder of the bogus rise is not clear.

In England some of the rises in the percentages of children gaining level 4 at the end of KS 2 may similarly be put down to teaching test technique and teaching to the test. This is in addition to the evidence discussed above which shows that the statutory procedures themselves were resulting in level 4 being given for lower quality work in English in 1999 compared with 1996. Further, the changing use of the anchor tests and the different patterns associated with Phase 1 and 2 suggest that the standard setting procedures have had an important impact on the published results.

Summary

Four separate perspectives have been used to look at the appropriateness of statutory test scores as a basis for monitoring standards at the end of primary schools in England.

1. The results since 1995 have followed patterns that in themselves raise questions about their validity. They rose steadily, with one hiccup to 2000 and then became abruptly flat. These have been referred to in this paper as Phases 1 and 2. Further, the mathematics and English results hugged one another over the years in a surprising fashion.
2. Independent data suggest that the rises seen in English should be broken down into reading and writing. The rises in writing, so far as the limited available data indicate, seem to be accurately reflected by the reported levels. The gains in reading were lower. The data suggest that during Phase 1 reading rose by 3.8 standardized points. In other words, the proportion of pupils attaining a level 4 should have risen from 48% to 58%, corresponding to an effect size of 0.25, rather than 75%. In Phase 2 the independent data agree with the statutory data in that little or no change occurred.
3. In mathematics during phase 1 the test rises were closer to the statutory data although the average suggests a rise of 7.7 standardised points. This translates into a change from 44% getting a level 4 to 64%, corresponding to an effect size of 0.51, rather than 72%. In Phase 2 the slight rise in statutory results is confirmed by the independent data although there is a suggestion that the actual rise might have been greater.
4. The rises in Phase 1 indicated above may be due, at least in part, to the children becoming more adept at taking tests as schools taught test technique.
5. The use of new tests every year with cut-scores to define levels severely restricts the use of the tests as a tool to monitor standards because: (a) there is an inherent limit to the accuracy with which the standards can be measured in any one year, and (b) cut-scores limit the extent to which standards can be examined over the ability

range even when cut-scores are reported for several levels. Analysis of the way in which cut-scores were set by the SCAA/QCA suggests that the shift from equating standards only to the previous year to maintaining standards over several years happened in 2000/01 and largely accounts for the different pattern of results in Phases 1 and 2.

6. Parallels from Texas suggest that similar things have been happening in that state. The apparently miraculous rise in high stakes test scores was found to be largely illusory in reading although there had been rises in mathematics. It seems that teaching test technique and teaching to the test may account for the rise.

Conclusion

Nearly a decade of national testing has generated a vast amount of data, which have been used for multiple purposes. This article has focused on one of its uses—monitoring standards over time—and in this national testing has failed for a number of reasons. The major points are list below:

1. The statistical procedures were faulty on one major feature, outlined above, which was not corrected until 2000/01.
2. The test data are used in a very high-stakes fashion and the pressure created makes it hard to interpret the data. Teaching test technique must surely have contributed to some of the rise, as must teaching to the test.
3. The official results deal with whole areas of the curriculum (English and mathematics) but the data suggest that standards have changed differently in different sub-areas. Writing improved much more than reading. The independent mathematics tests showed different patterns from one another, as did different questions within Brown *et al.*'s (2003) study. No data are available on changes in vocabulary levels and so on.
4. The form of the national tests has changed over time. It is therefore very hard to answer questions such as: to what extent can rises be attributed to the tests becoming more pupil-friendly?
5. The curriculum itself inevitably evolves and when it does the content of the tests must follow. This makes the task of monitoring standards using statutory tests particularly problematic.
6. There is always a concern about independence. There is little doubt that the SCAA/QCA have acted independently in setting standards but they have very close links with the DfES and there must always be a concern in the public mind about decision-making.

Statutory test data must not be used to monitor standards over time. We need a new independent body dedicated to the task of monitoring over time. The body would need time and money to set up a system and this need not be expensive. It could build on the lessons learned by similar bodies such as the APU in England, the NAEP in the USA, and the Assessment of Achievement Programme in Scotland established in

1981 by the Scottish Office Education and Industry Department. The new body would, in the long run, generate data of considerable educational importance.

Acknowledgements

Thanks are due to Brian Henderson of the CEM Centre, who was responsible for extracting the PIPS data and Nicola Foster, also from CEM, who provided the MIDYIS data.

Several people offered helpful comments on an initial draft of this article, which was presented at the BERA meeting in Edinburgh in 2003, and they were much appreciated. They include: Dr Paul Newton of the QCA, Dr Ian Schagen of the NFER, Dr Linda Croxford of the University of Edinburgh and Dr Steven Strand, NFER-Nelson. Finally, thanks are due to the two anonymous referees for their helpful comments.

References

Barber, M. (2001) The very big picture, *School Effectiveness and School Improvement,* 12(2), 213–228.

Brown, M., Askew, M., Rhodes, V., Denvir, H., Ranson, E. & Wiliam, D. (2003) Characterising individual and cohort progression in learning numeracy: results from the Leverhulme 5-year longitudinal study, paper presented at the *American Educational Research Association annual meeting*, Chicago, IL, April.

Davies, J. & Brember, I. (1997) Monitoring reading standards in Year 6: a seven year cross-sectional study, *British Educational Research Journal,* 23(5), 615–622.

Davies, J. & Brember, I. (1999) Standards in mathematics in Years 2 and 6: a nine year cross-sectional study, *Educational Review,* 51(3), 243.

Davies, J. & Brember, I. (2001) A decade of change: monitoring reading and mathematics attainment in Year 6 over the first ten years of the Education Reform Act, *Research in Education,* 65, 31–40.

Fitz-Gibbon, C. T. (1992) Peer and cross-age tutoring, in M. C. Alkin (Ed.) *Encyclopedia of educational research* (New York, Macmillan), 980–984.

Fitz-Gibbon, C. T. (1996) *Monitoring education: indicators, quality and effectiveness* (London, Cassell).

Jensen, A. (1980) *Bias in mental testing* (London, Methuen).

Hilton, M. (2001) Are the Key Stage Two reading tests becoming easier each year? *Reading*, April, 4–11.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F. & Stecher, B. M. (2000) *What do test scores in Texas tell us*? (Santa Monica, CA, Rand).

Massey, A., Green, S., Dexter, T. & Hammet, L. (2003) *Comparability of national tests over time: KS1, KS2 and KS3 standards between 1996 and 2001. Final Report to QCA of the Comparability Over Time Project* (Research and Evaluation Division of the University of Cambridge Local Examinations Syndicate).

Minnis, M. & Higgs, S. (2001) *Evaluation of the National Literacy and Numeracy Strategies: technical report for the Testing Programme 1999–2001*. Available online at: www.qca.org.uk

National Audit Office (2001) *Modern policy-making: ensuring policies deliver value for money*. Report by the Comptroller and the Auditor General London, House of Commons, Stationery Office.

Quinlan, M. & Scharaschkin, A. (1999) National curriculum testing: problems and practicalities, paper presented at the *British Educational Research Association Annual Conference*, Brighton, September.

Ruddock, G. (2000) *Third International Mathematics and Science Study Repeat (TIMSS-R): First National Report*. Brief RB234 (London, Department for Education and Employment).

Teddlie, C. & Reynolds, D. (Eds) (2000) *The international handbook of school effectiveness research* (London, Falmer Press).

Tymms, P. (1999) *Baseline assessment and monitoring in primary schools: achievements, attitudes and value-added indicators* (London, David Fulton).

Tymms, P. & Albone, S. (2002) Performance indicators in primary schools, in: A. J. Visscher & R. Coe (Eds) *School improvement through performance feedback* (Lisse, Swetz & Zeitlinger), 191–218.

Tymms, P. & Coe, R. (2003) Are standards rising in English primary schools?, *British Educational Research Journal,* 29, 639–653.

Tymms, P. & Fitz-Gibbon, C. T. (2001) Standards, achievement and educational performance, 1976–2001: a cause for celebration? in: R. Phillips & J. Furlong (Eds) *Education, reform and the state: politics, policy and practice, 1976–2001* (London, Routledge).